NAVAL RESEARCH LOGISTICS QUARTERLY

SEPTEMBER 1976
VOL. 23, NO. 3

OFFICE OF NAVAL RESEARCH

NAVSO P-1278

NAVAL RESEARCH LOGISTICS QUARTERLY

EDITORS

Murray A. Geisler
Logistics Management Institute

W. H. Marlow
The George Washington University

Bruce J. McDonald
Office of Naval Research

MANAGING EDITOR

Seymour M. Selig
Office of Naval Research
Arlington, Virginia 22217

ASSOCIATE EDITORS

Marvin Denicoff
Office of Naval Research

Alan J. Hoffman
IBM Corporation

Neal D. Glassman
Office of Naval Research

Jack Laderman
Bronx, New York

Thomas L. Saaty
University of Pennsylvania

Henry Solomon
The George Washington University

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logist
will publish research and expository papers, including those in certain areas of mathematics, statistics, and ecor
relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March
September, and December and can be purchased from the Superintendent of Documents, U.S. Government Pr
Office, Washington, D.C. 20402. Subscription Price: $11.15 a year in the U.S. and Canada, $13.95 elsewhere. (
individual issues may be obtained from the Superintendent of Documents.

The views and opinions expressed in this Journal are those of the authors and not necessarily those of the
of Naval Research.

Issuance of this periodical approved in accordance with Department of the Navy Publications and Printing Regul
P-35 (Revised 1-74).

A SURVEY OF MAINTENANCE MODELS: THE CONTROL AND SURVEILLANCE OF DETERIORATING SYSTEMS*

William P. Pierskalla and John A. Voelker

Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, Illinois

ABSTRACT

The literature on maintenance models is surveyed. The focus is on work appearing since the 1965 survey, "Maintenance Policies for Stochastically Failing Equipment: A Survey" by John McCall and the 1965 book, *The Mathematical Theory of Reliability*, by Richard Barlow and Frank Proschan. The survey includes models which involve an optimal decision to procure, inspect, and repair and/or replace a unit subject to deterioration in service.

INTRODUCTION

For nearly two decades, there has been a large and continuing interest in the study of maintenance models for items with stochastic failure. This interest has its roots in many military and industrial applications. Lately, however, new applications have arisen in such areas as health, ecology, and the environment. Although it is not possible to detail these many applications of maintenance models, some of them are: the maintenance of complex electronic and/or mechanical equipment, maintenance of the human body, inspection and control of pollutants in the environment, and maintenance of ecological balance in populations of plants and animals.

Just as the interest in maintainability has grown and changed, so has the sophistication of the models and control policies for solving the maintenance problems. Many of the important early maintenance and/or inspection models possessed an elegance and simplicity which led to easily implementable policies. These models were later generalized; and although much of the elegance and simplicity remains, some of the new results are complex and require the use of large computers for implementation.

In two earlier and excellent works, the area of maintainability was rather thoroughly researched and surveyed up to 1965. These works are: *The Mathematical Theory of Reliability*, by Richard Barlow and Frank Proschan [1965] and the paper, "Maintenance Policies for Stochastically Failing Equipment: A Survey," by John McCall [1965]. Here, the presentations of these earlier works will not be repeated, but rather the coverage is primarily the results which have appeared since 1965. There are a few exceptions to this rule, however, when it was necessary to review some key models which appeared prior

*This survey was produced in part as a result of research sponsored by the Office of Naval Research under Contract N00014-67-A-0356-0030, Task NR042-322. Reproduction in whole or in part is permitted for any purpose of the United States Government.

to 1965 and which formed the foundation of many of the later studies. Furthermore, in this survey it assumed the reader is familiar with such concepts as Markov processes, Poisson processes, dynam programming, linear programming, Lagrange multipliers, and other concepts and techniques with whic the holder of a master's degree or an experienced practitioner in statistics or operations research woul be familiar.

Since there are many threads of activity which contribute to the total fabric of maintainability an since it is not possible to include all of them in a reasonably sized survey paper, this study include only those models which involve an optimal *decision* to procure, inspect, repair and/or replace a un subject to deterioration in service. Not included are models which describe the operating characteristi of a system, such as repairmen or machine interference models, unless these models involve some opt mization as well. Also not included are models which deal with the controlled reliability of a system suc as the design of redundant systems unless they also include aspects of optimal inspection, repair, and/ replacement decisions. Each of these excluded areas is large, important, and worthy of a survey i its own right. Furthermore, each lies slightly outside of what many people would call maintainabilit Even with these restrictions, however, the reader will see that the many topics covered in this survey a rich with theoretical significance and practical application.

There are many possible ways to classify the works in maintainability. One could establish a multi dimensional grid whose coordinates would be (i) states of the system, such as deterioration level, ag number of spares, number of units in service, number of state variables, etc., (ii) actions available, suc as repair, replacement, opportunistic replacement, replacement of spares, continuous monitorin discrete inspections, destructive inspections, etc., (ii) the time horizon involved, such as finite or infini and discrete or continuous, (iv) knowledge of the system, such as complete knowledge or partial know edge involving such things as noisy observation of the states, unknown costs, unknown failure distrib tions, etc., (v) stochastic or deterministic models, (vi) objectives of the system, such as minimize lo run expected average costs-per-unit time, minimize expected total discounted costs, minimize tot costs, etc., and (vii) methods of solution, such as linear programming. dynamic programming, generalize Lagrange multipliers, etc.

Then in each cell of the grid, one could conceivably place every paper written on maintainabilit The empty cells would mean those research areas were irrelevant or not yet explored. To some exte McCall [1965] uses such a classification scheme (on a modest scale) by considering the primary catego ies of "preparedness" and "preventive maintenance" models with and without complete informatio

Although classification schemes based on some breakdown involving two or more of the sev categories mentioned above are useful for establishing an underlying general theory of maintainabili and also for exposing some unexplored areas, such a scheme has not been rigidly followed here. Rath the papers have been classified in such a manner that a practitioner might be able to find the mode relevant to his maintenance problem (to a large extent this approach parallels that taken by Barlow a Proschan [1965]. Thus, there are two major sections with several subsections in each.

The first section surveys discrete time maintenance models. That is, at discrete points in time, a un (or units) is monitored and a decision is made to repair, replace, and/or restock the unit(s). The secor section surveys continuous time maintenance models. Actions and events are not *a priori* restricted occur only within a discrete subset of the time axis.

In the discrete time section, most of the models are Markov decision models in which the state

e system is described by the level of deterioration and/or the number of spare units available in in-ntory. This section is subdivided by segregating models with no restocking from those which involve inventory (restocking) decision. The subsection on models with no restocking is further subdivided the basis of kinds of information available.

The continuous time models are subdivided into several topic categories also. The first of these is ncerned with applications of control theory to maintainability. The control function $m(t)$ acquires e interpretation "rate of maintenance expenditure at time t." By the use of control theory, $m(t)$ is lected to maximize discounted return. The second subsection deals with age replacement models reby giving new enhancements to the classical model, such as age dependent operating cost and the timal provision of spares. Also described in this subsection are certain policies that regulate time of placement according to the occurrence of the kth failure. The third subsection describes shock models, ., systems where the unit is subject to external shocks according to some stochastic process, and se shocks affect its failure characteristics. Subsection four deals with interacting repair activities. re the decisionmaker must control a system composed of more than one unit. Specifically discussed such "system-wide" activities as opportunistic replacement, cannibalization, multistage replace-nt, and variable repair rate. Also discussed are models involving incomplete information. The cisionmaker's information may be incomplete only in regard to the current state of the system, in ich case inspection models arise. On the other hand, he may not know completely the probability governing the system or the actual cost implication of various actions. In the former case, max-min Bayesian strategies have been devised. In the latter case, the statistical implications of a related mpling procedure are presented. The concluding section indicates some areas for future research.

Although the authors have tried to give a reasonably complete survey, the reader will note that ranslated Eastern European and Asiatic papers are missing. Any other papers that are not included re either inadvertently overlooked, not directly bearing on the topics of this survey, or no longer in nt. The authors apologize to both the readers and the researchers if we have omitted any relevant pers.

1 DISCRETE TIME MAINTENANCE MODELS

In this section, models are reviewed which utilize information regarding the degree of deteriora-ti of the unit or units in order to select the best action at certain discrete points in time. The deteri-o tion may be described by such factors as "wear and tear," "fatigue," etc. In some cases, inspection d isions must be made in order to ascertain the current state of the unit before a repair or replacement a on is taken. In others, it may be assumed that the current state is always known at the beginning of a eriod and the available actions are to replace the unit or to choose one of several repair activities w ch will tend to decrease the degree of deterioration. Often the decisionmaker must take these actions u er conditions of incomplete information about costs, underlying failure laws, or noisy observations of t unit's state. Models involving complete knowledge are treated first, then the models with incomplete in rmation are covered. In these latter models, it is always assumed that there are new replacements a ilable (i.e., an unlimited supply of new spares). The section concludes with models involving the p odic restocking of inventories of spare parts.

Because of the Markovian nature of the state and action transition processes, Markov decision th ry and inventory theory are the primary approaches used in the model formulations. Consequently,

linear programming and dynamic programming are the primary techniques used to solve these mod

COMPLETE INFORMATION

A unit (or several units) is inspected every period and a decision is made to repair or replace unit whenever it is found to be in a certain set of states. In the absence of a decision to repair or repla it is assumed that the unit deteriorates stochastically through a finite set of states denoted by the of integers $\{0, 1, \ldots, L\}$ according to a Markov chain. The state 0 denotes a new or complet renovated unit and the state L an inoperative or failed unit. After inspecting the unit a decision is m either to repair, replace, or do nothing to the units. In most of the models there are only two decisi every period: decision 1 means do nothing and decision 2 means replace.

Depending upon the assumptions concerning the time horizon, the amount of information av able, the nature of the cost functions, the objectives of the models, system constraints, and numb of units, different authors have produced many interesting and significant results for variations on t basic model. The basic model was originally introduced by Derman [1962] and extended by Klein [19 Although these two important papers are clearly presented in Barlow and Proschan [1965] and Dern [1970], the basic model is briefly given here prior to surveying later generalizations.

The unit is observed at times $t=0, 1, 2, \ldots$ to be in one of the states $X_t \in \{0, 1, \ldots, L\}$. If action (decision 1) is taken, then p_{ij} denotes the probability of moving from state i to state j in period. If the unit is replaced (decision 2), then the unit moves immediately into state 0, and the tra tion during the period is governed by the probabilities $\{p_{0j}\}$. It is assumed that

(1) $$p_{i0} = 0, \qquad i = 0, \ldots, L-1,$$

(2) $$p_{L0} = 1, \quad \text{and}$$

(3) $$p_{iL}^{(t)} > 0 \qquad \text{for some } t \text{ and each } i = 0, \ldots, L-1.$$

Condition (1) implies the unit is never as good as "new" after its first period of service; condi (2) implies the unit must be replaced on failure; and condition (3) implies the unit will eventually and that the underlying Markov chain, $\{X_t\}$, has a single ergodic class and the steady state probabili exist.

Upon inspecting the unit at any time, it is possible to replace the item before failure. In this it may be possible to avoid the consequences of failure or further deterioration of the unit. These cisions to replace the unit or do nothing are summarized by a set of decision rules (or replacem rules) based on the entire history of the process up to time t. Although the most generality is achie by considering these rules as elements of the class of nonstationary, randomized rules, conditions h been established by Derman [1970] which enable one to restrict attention to stationary nonrandomi rules. These latter rules are of the type that the set $\{0, 1, \ldots, L\}$ is partitioned into two subset and $\mathscr{D}$. If $X_t \in \mathscr{R}$ then replace the unit and if $X_t \in \mathscr{D}$, do nothing. Most of the models surveyed here sat the conditions given by Derman or else restrict their attention only to the class of stationary nonr domized rules which do not utilize all prior history and can be denoted by R_i for $i = 0, \ldots, L$, wh R_i is the action taken when $X_t = i$.

By making an intervening decision before observing state L, the behavior of the system is modifi and the evolution of the system under a replacement rule results in a modified Markov chain. T

y costs consist of c_1 to replace a unit that has not yet failed, and a higher cost c_2 to replace a failed t.

The objective is to minimize the expected long run average cost-per-unit time.

By making an additional assumption on the original transition probabilities

$$\sum_{j=k}^{L} p_{ij} \text{ is nondecreasing in } i = 0, \ldots, L-1$$

$$\text{for each fixed } k = 0, \ldots, L.$$

rman has shown that the optimal replacement rule R^* is a "control limit" rule; that is, there is a te $i^* \in \{0, 1, \ldots, L\}$ such that if the observed state k satisfies $k \geqq i^*$ then replace the unit and if k i^* do nothing. This key result reduces the size of the set of rules R in the minimization from 2^{L-1} es to at most $L+1$ rules. The same key result holds when the objective is changed to minimize the al long run discounted costs.

Assumption (4) implies that if no replacements are made, the probability of deterioration increases the initial state increases. Barlow and Proschan [1965] point out that this conditional probability is quivalent to assuming the original Markov chain is IFR.

The mathematical programming problems resulting from the above objectives have a natural formuon as dynamic programs, and in this context, successive approximations and policy iteration techues may be used (see Bellman [1957], Bellman and Dreyfus [1962], and Howard [1971]). Derman provided an interesting formulation in the long run average cost case leading to a linear programming blem. On the surface it may appear that the linear programming problem is very difficult to handle most important problems because of its large size; however, recent work in large-scale linear proamming (cf. Lasdon [1974]) gives hope for solving these very large Markov decision problems.

Kolesar [1966] considers the same model with the cost function generalized to allow an "occucy" cost, A_i, associated with being in each state i. With the added condition $0 \leqq A_1 \leqq \ldots \leqq A_L$, demonstrates that a control limit rule is optimal in the long run average cost case.

Now the $L+1$ control limit rules can be represented by the integers $\{i | i = 0, \ldots, L\}$. In his case ere a control limit rule (or rules, if uniqueness is not present) is optimal, Kolesar shows that the cost ction is integer quasi-convex in i, i.e., if i^* is the smallest optimal control limit then φ_{R_i} is nonreasing for $i < i^*$ and nondecreasing for $i \geqq i^*$ (see Greenberg and Pierskalla [1971] for results on si-convexity). This observation enables him to develop more efficient algorithms to solve the average c problem. He further observes that in the linear programming formulation of the model the rection to control limit rules allows one to modify the simplex algorithm so that at most L changes of sible bases are needed before the optimal rule is obtained.

The results of Derman and Kolesar were later extended by Ross [1969a] to the case where the state he system can be represented by an element of some nonempty Borel subset S of the nonnegative line. Thus, continuous or denumerable state spaces are now allowed. Let $x \in S$ denote the state of system upon inspection at the beginning of a period, and F_x the cumulative distribution function cribing the next state given that the decision is made not to replace defined for $x > 0$. The failed state enoted by 0. Thus $p(x) \equiv F_x(0)$ is the probability the unit currently in state x will fail before the next ection.

In parallel with the finite state models, a control limit policy, R_y, replaces the unit at time t if X_t or $X_t > y$ for some control limit $y \in (0, \infty]$ and does nothing otherwise. Ross then proves the following $g(x)$, the cost of being in state x, is a nondecreasing bounded function of x for $x > 0$, p_x is a nondecr ing function of x for $x > 0$, and for each $y > 0$ the function $(1 - F_x(y))/(1 - p_x)$ is a nondecrea function of x for $x > 0$, then there is a control limit policy R_y that is optimal in the long run discour case. Under some additional conditions on $g(x)$ and $F_x(y)$, he then shows that a control limit polic optimal for the long run average cost case as well.

In a recent paper Kao [1973] considers a different type of generalization of the basic Derman mo He considers the problem of a deterioration process which moves from state 0 to state L, but the t spent in each state before a transition is a random variable depending on the transition. This prob is modelled as a discrete time finite state semi-Markov process. Since the underlying process is on deterioration it is assumed that $p_{ij} = 0$ for $j \leqq i$ and that the holding time in state i before going to a positive, finite, integer-valued random variable. As in Kolesar, Kao assumes an occupancy c A_i per unit time, a fixed cost c_i of replacement and a variable cost v_i of replacement per unit time decision to replace is made when the unit is in state i. Under reasonable conditions on the cost functi and the expected sojourn time in each state, Kao proves that a control limit policy is optimal over class of stationary nonrandomized policies for the long run average cost case. He then generalizes model to include the time the unit has spent in the state as well as the state itself in determining wh to replace the item.

Until now, it has been assumed that the decisionmaker is continuously aware of the state of unit. That assumption will now be dropped. To ascertain the true state of the system, the decisionma must inspect it, which is an action entailing a particular cost. Rules for scheduling inspections now en into the policy.

In an early paper, Klein [1962] was concerned with the different levels of decisions which co be made upon the inspection of a unit. These decisions ranged from replace with a new unit to vari types and degrees of repair. In addition, he considered the decision as to when the next inspect should be taken. He showed that the problem could be formulated as a linear fractional program a then reduced to a linear program to be solved.

More specifically, there are costs associated with inspections and repairs. A repair moves the u from state j to state i at cost $r(j, i)$; replacement is equivalent to moving it from state j to state 0. T cost of inspection while in state j is $c(j)$. After an inspection, the decisionmaker can choose to s the next m time periods before making another inspection. It is assumed that m is bounded by so number, say M. If the unit is discovered to have failed at some time before inspection, it is assumed t its failure time can be determined during the inspection.

The states are $\{0, 1, \ldots, L, L(1), \ldots, L(M)\}$, where $L(m)$ denotes that the unit failed m ti periods ago. The action (j, m) denotes placing the unit in state j by repair or replacement and skippin periods until the next inspection. The transition probabilities are the same as before. For this mode is assumed $p_{ij} = 0$ if $i > j$, $p_{L, L(1)} = p_{L(1), L(2)} = \ldots = p_{L(M), L(M)} = 1$, and $p_{iL}^{(t)} > 0$ for some t and $i = 0, \ldots, L-1$.

A linear programming formulation for the long run average cost case can be found in Klein. Derm [1970] shows that a stationary policy is optimal.

Much in the spirit of Derman's and Klein's models, but by using the functional equation of dynan programming, Eppen [1965] gives sufficient conditions on the cost functions and transition matric

show that the total discounted expected costs are minimized from period n to the end of the finite rizon by the following policy: there exists a sequence $\{i_n^*\}$ of critical numbers such that in period n he state of the deteriorating unit exceeds i_n^*, then return the system to state i_n^* (by repair); otherwise not repair. This policy parallels the critical number policy in the single product stochastic inventory odel without setup cost.

In an interesting paper, Hinomoto [1971] considers the sequential control of N homogeneous units l generalizing the approach developed by Klein. At each decision point the decisionmaker determines e level of the repair activity and the time of the next inspection. Hinomoto considers two different ans for the maintenance of the N units. The first plan is a fixed rotating sequence for the units, i.e., d an optimal permutation schedule from the set of all permutations of the numbers 1, . . ., N and n inspect (and repair, if needed) the units sequentially in the order given by the permutation. The ond plan is a priority plan which alters the fixed sequencing by queueing those units for repair or placement which have fallen below a certain critical level of performance. (In the first plan, such a it would just have to wait its turn in the permutation sequence.) He formulates both problems as ear programs and finds the optimal decisions over the class of fixed cyclic inspection patterns (ex- ot for the temporary displacing priorities in the second plan).

Other authors Derman [1970], Eckles [1968], Sondik [1971], and Smallwood and Sondik [1973] a allow for this greater variety in the decision space by allowing repair or replacement decisions j 0, 1, . . ., L} (more will be said about their results in the next section).

omplete Information

A problem which often arises in the study of maintenance is the lack of information concerning ny aspects of the model. This lack of information may take many different forms. In his earlier s vey, McCall [1965] treated the case of lack of knowledge of the underlying failure distribution. In a lition to this case, there can be lack of information due to (i) random costs or unknown costs and i noisy observations on the state of the system. A paper treating the costs as random variables with k wn distribution is given by Kalymon [1972]. The papers concerning unknown costs are generaliza- ti s by Kolesar [1967] and Beja [1969] of another basic model due to Derman [1963a]. The papers on n sy observations are by Eckles [1968], Sondik [1971], and Smallwood and Sondik [1973]. Concluding th section, the papers by Satia [1968] and Satia and Lave [1973] on the lack of knowledge of the u erlying probability distribution are mentioned.

Kalymon [1972] generalizes the Derman model of the previous section by considering a stochastic re acement cost determined by a Markov chain. This chain is assumed to be conditionally independent of the Markov chain defining the deterioration of the unit from period to period. The cost, C, of a new u in period t is a random variable which takes on a finite set of values. There is a separable salvage va e of $-[r(c_t) + s(x_t)]$ and an occupancy cost $A(x_t)$ when the machine is in state x_t and the realized a ual cost of a new machine is c_t in period t.

When the cost functions $c + r(c)$, $s(x)$, and $A(x)$ are nondecreasing in their arguments, the Markov c ins $\{X_t | t = 0, 1, \ldots\}$ and $\{C_t | t = 0, 1, \ldots\}$ have distribution functions which are IFR, i.e., th y satisfy (4) and its analog, and the conditional expected costs satisfy a certain condition, then th optimal policy is a control limit policy for the nonstationary finite horizon discounted cost function. Th s result is then generalized to the infinite horizon ergodic chains case for both the long run dis-

counted and the long run average cost cases. In this context, there can be a different control limit pol for each realized replacement cost c.

In another interesting variation on his earlier basic model, Derman [1963a] considered the probl where the unit is in one of several operative states 0, 1, . . ., n and several inoperative states n-1, . . ., L. A decision $k \in \{1, \ldots, K\}$ is made to replace or partially repair the unit at some leve once it is observed in states 1, . . ., n and only to replace if in states $n+1, \ldots, L$. There is no expli cost structure available nor can one be easily inferred other than that the cost of failure is much grea than the cost of replacement. The objective is to maximize the expected length of time between repla ments, called the cycle time, subject to the constraint that the probability a replacement is made wh the process is in state j is no greater than some preassigned number α_j for $j=n+1, \ldots, L$. Derm develops a linear programming formulation of this model over the class of rules which repeat eve time the unit is renewed.

In a subsequent paper, Kolesar [1967] restricts the set of decisions to only two: replace or do n replace. From the class of stationary randomized rules, a generalized control limit rule is defined by

$$\begin{aligned} &\text{do not replace} && \text{when } i < m, \\ &\text{replace with probability } \frac{\beta\pi_i(R_m)}{\beta\pi_i(R_m)+(1-\beta)\pi_i(R_{m+1})} && \text{when } i = m, \\ &\text{replace} && \text{when } i > m, \end{aligned}$$

where $0 \leqq \beta \leqq 1$ and R_m and R_{m+1} are control limit rules (as defined previously). The coefficien which forms the convex combination of the steady state probabilities for the two nonrandomized cont limit rules R_m and R_{m+1} is obtained as the coefficient which satisfies $\pi_j^* = \beta\pi_j(R_m) + (1-\beta)\pi_j(R_m$ and the π_j^* minimize π_0 on the set $\sum_{j=0}^{L} \pi_j = 1$, $\sum_{i=0}^{L} \pi_i p_{ij} = \pi_j$, $\pi_j \geqq 0$ and $\frac{\pi_L}{\pi_0} \leqq \tau$, where τ is the p determined maximum tolerable probability of failure in a cycle. Under the IFR assumption (4), Koles establishes that the optimal stationary policy is either do not replace until state L, or else there a control limit rules R_m and R_{m+1} such that a generalized control limit rule is optimal when the maximu tolerable probability of failure in a cycle is τ.

Beja [1969] does not make the IFR type of assumption (4) on the transition probabilities. In so doi he is able to consider more general transition matrices, e.g., the "bath-tub" variety. Let y_i and x_i deno the probability of failure before replacement and the expected time before replacement given th the present state is i, respectively. For every state i the set of constraints

$$y_i/x_i \leqq \tau, \qquad \text{for a given } \tau,$$

impose restrictions on the potential hazard encountered while in state i. With these additional co straints on the unit, Beja shows that from the class of stationary randomized policies, one of the 2^L nonrandomized policies (i.e., replace or keep whenever the process is in state 1, 2, . . ., $L-1$) optimal. He then demonstrates what the "implicit" replacement and failure costs are, given the optim solution which maximizes the cycle length.

The problems of uncertainty in studying a maintenance problem are treated somewhat differently Eckles [1968], Sondik [1971], and Smallwood and Sondik [1973]. The underlying process for the it is still the finite state, discrete time Markov chain as before and there are K possible decisions m which a selection k is to be made. For example, these decisions could be to replace, repair or do thing. Let $p_{ij}^k(t)$ be the true conditional probability that the unit goes to state j in the next period en that it is in state i at period t and decision $k \equiv k(t)$ was taken. An inspection of the unit after taking cision k in state i during period t yields the observation x where $q_{xi}^k(t)$ is the conditional probability outcome x given state i and decision k. In this manner, it is possible to characterize maintenance-pection problems ranging from those whose observation of the state reveal knowledge of the actual te with certainty (as in the previous models) to those which provide no information of the new state the system. It is assumed (i) that the actual age of the unit in use in period t is always known with tainty given the sample history of observations and decisions $\mathcal{H}_t$, and (ii) there is an underlying cost $t)$ which represents the cost of going to state j, given state i and decision k in period t. Letting

$$c_i^k(t) = \sum_{j=0}^{L} p_{ij}^k(t) c_{ij}^k(t)$$

the expected one period cost given (i, k, t), and assuming that for any sample history $\mathcal{H}_t$ the one-p transition matrix $P(\mathcal{H}_t)$ and the current age of the unit $t(\mathcal{H}_t)$ are jointly a sufficient statistic for , then there exists an optimal solution which minimizes the expected total discounted costs where only information needed are the one-step transition matrix $P(\mathcal{H}_t)$ and the current age of the unit $t)$ and not the entire prior history $\mathcal{H}_t$. Bayes Theorem is then used to update $P(\mathcal{H}_t)$ to $P(\mathcal{H}_{t+1})$ $P(\mathcal{H}_{t-1})$ if one numbers backward as is often done in finite time dynamic programming). Eckles mulates the problem as a dynamic program and under the standard assumption that whenever a t is replaced it is completely renewed (its age is 0) and its transition probabilities are then inde-ident of the past history of the process, he presents an algorithm for finding an optimal nonran-nized age replacement policy for this renewal process.

Sondik [1971] and Smallwood and Sondik [1973] treat the same problem as Eckles; however, y demonstrate that the entire history of the process is contained in the information vector

$$\pi(t) = (\pi_0(t), \ldots, \pi_L(t)),$$

ere

$\pi_j(t) \equiv$ conditional probability that the actual state at time t is j given observation x, history $\mathcal{H}_{t-1}$ and decision $k(t)$,

that by Bayes Theorem

$$\pi_j(t+1) = \frac{\Sigma_i \pi_i(t) p_{ij}^k(t) q_{xj}^k(t)}{\Sigma_i \Sigma_j \pi_i(t) p_{ij}^k(t) q_{xj}^k(t)}.$$

Thus, if $\hat{\pi}(t)$ is the statistic generated by the outcomes $\mathcal{H}_t$ then $\hat{\pi}(t+1)$ is a sufficient statistic for $\mathcal{H}_t$. This approach differs slightly from the Eckles model in that

$q^k_{xj}(t) \equiv$ conditional probability that the outcome x is observed given that the true outcome is j and that decision k was taken just prior to the inspection.

Thus, $\pi(t)$ behaves as a discrete time, continuous state Markov process. They then demonstrate th the expected total discounted cost function is piecewise linear and convex. Using this structure th develop an efficient algorithm which makes the rather difficult large state space problem relativ easy to solve.

In conclusion of this subsection the case of incomplete knowledge of the probability law governi the system's evolution is briefly considered. As was mentioned earlier McCall [1965] surveyed this ca extensively.

The optimization of a maintenance problem modelled as a Markov chain with unknown transit probabilities is usually approached from either a game theoretic (i.e., max-min or max-max) or from Bayesian point of view. The max-min approach essentially seeks the maximum overall possible polic of the minimum of the total expected long run discounted return with respect to all possible transiti matrices (within some set). The max-max approach is defined analogously. Satia [1968] proves the exists a pure, stationary policy for this latter decision process that is optimal.

Satia and Lave [1973] discuss this earlier Satia result and present an algorithm similar to Howar Policy Improvement Algorithm. Their algorithm will converge to within any predetermined ϵ-inter about the true max-min solution in a finite number of iterations. Satia and Lave also formulate a B esian approach to the problem and present an implicit enumeration algorithm for its solution.

Maintenance and Inventory Models

Most maintainability models which provide for the replacement of a unit assume that the repla ment items are drawn from an infinite stock. However, for some models, this stock is not infinite; inde its management becomes a control variable.

There are many inventory papers which treat stock replenishment problems for stochastically fa ing equipment: Falkner [1969], Prawda and Wright [1972], Sherbrooke [1968, 1971], Sobel [1967], P teus and Lansdowne [1974], Silver [1972], Miller [1973], Moore, et al. [1970], Demmy [1974], and Drir water and Hastings [1967]. These papers do not consider problems where a decision must be made repair, replace or inspect a unit or units. Rather they assume that once a unit has failed, it must be r paired or replaced and the decision process is how much inventory to stock either initially, periodical or continuously.

To be more specific, Prawda and Wright [1972] examine a system where there may be many ide tical units in operation. These units fail in one of two ways. The unit fails and is repairable with pr ability distribution $F_i(\cdot)$ or nonrepairable with probability distribution $\mathcal{R}_i(\cdot)$ in period $i = 1, 2, \ldots$ R pairable units go to a repair facility and after $\mathcal{N}$ periods are renewed and returned to inventory. N repairable units are discarded. If the available inventory is insufficient to meet replacement nee these needs are backlogged until inventory is available. An order for new units is placed at the beg ning of each period and delivery is received λ periods later. They consider the two problems of h much to order every period to minimize the expected total discounted costs and to minimize the

ted long run average cost-per-unit time. In their model, there are four costs: ordering, holding, rtage, and salvage. Following some earlier work in inventory theory by Veinott [1965], Prawda and ight show that for an ordering cost of c per unit and explicit quasi-convexity of the single period t function, the optimal policy is a stationary single critical number in each period. That is, in period id before ordering in this period if the state of the system determined by the stock on hand, in repair, l on backorder is denoted by x_t, then there is a number $\bar{y}$ such that

$$\text{if } x_t < \bar{y} \text{ order } \bar{y} - x_t,$$

erwise do not order. They also consider the case where there is a setup cost K every time an order laced and give results on the optimal order quantities.

Taking a different approach, Sherbrooke [1968, 1971] considers the problem of determining the k levels at each echelon of a multiechelon multiunit inventory system of repairable units. The idea iat at various sites $j = 1, \ldots, J$ there are repairable units of types $i = 1, \ldots, I$ in use and in inven- , y_{ij}; and there is a central facility, 0, which maintains a buffer inventory, y_{i0}, for use at the sites n needed. Each site including the central facility has repair capabilities. The problem is to determine y_{ij} for $i = 1, \ldots, I$ and $j = 0, 1, \ldots, J$ which minimize the total weighted expected number of units klogged at any point in time subject to a budget constraint on repair and operating costs. The failure nits are independent and identically distributed (by unit type) according to a logarithmic Poisson ribution with constant variance to mean ratios. There is no transshipment among sites $1, \ldots, J$ the repair decisions are not explicitly entered in the model (only implicitly through the logarithmic son failure rates). It is also assumed that there are an infinite number of repair activities, so that repair times are independent of the number of units being repaired. Sherbrooke shows that the gen- ized Lagrange multiplier approach of Everett [1963] and Greenberg and Robbins [1972] can be used otain near optimal solutions to the problem.

Porteus and Lansdowne [1974] consider the same model, but assume the failure process is Poisson, $y_{i0} = 0$ for all $i = 1, \ldots, I$ and that the mean repair times for the repair of unit type i at site j under- g repair work of level k can be controlled. Using a generalized Lagrange multiplier algorithm, they in the spare stock quantities and mean repair times which minimize either the long run average -per-unit time or the long run total discounted cost.

In a related paper also based on Sherbrooke's work, Silver [1972] determines the inventory levels, r repairable subassemblies ($i = 1, \ldots, I$) of a major assembly. The failure process is Poisson for subassembly, the failure of a single subassembly makes the entire unit inoperative, and canni- ation of major assemblies awaiting repair is assumed. Under the additional assumption that there inventory of the major assembly (i.e., $y_0 = 0$), Silver shows that the entire optimization problem is rable and easily solved. In the case $y_0 > 0$, he gives a near-optimal algorithm to obtain y_i for $i = 0, \ldots, I$.

Miller [1971] considers the optimal stocking problem also; however, in [1973] he asks the question nich site should a repaired item be sent after it is repaired at a central facility given that transpor- n times differ and that only the central facility can repair items. He formulates a single-item multi- ion model and shows by simulation that the policy of shipping the repaired item to the site which zes the greatest marginal decrease in the expected backorders from this one additional unit (com- d over the time required to transport the unit) is better than the current Air Force shipping policy.

Furthermore, under the assumption that repairs are instantaneous after a failure, he proves that th policy is optimal.

Along different lines, Derman and Lieberman [1967] consider a joint replacement and stockin problem. Inspections are made every time period. An initial stock of N identical units are on hand. the end of each period, a decision is made to replace the currently working unit or not to replace it. a replacement is made, the new unit works at a level s with probability f_s and continues to work at th same level until it either fails or is replaced. Its life-length is a random variable with a geometric di tribution. If at the end of the period the unit in service has failed, it is replaced provided there are uni still in stock. If there are none in stock, the system is down for one period of time while N units a reordered.

The state space is described by $\{(n, s): n=1, \ldots, N; s=1, 2, \ldots\} \cup \{0\}$, where n denotes th number of (identical) units on hand including the one in service, and s denotes the performance level the unit in service. Note that the levels of service are a denumerable set. The element 0 denotes n units in stock and the system is down. The available actions are $\{1, 2\}$, where 1 denotes no replaceme and 2 denotes replacement (or reorder).

Using the resulting transition probabilities and the cost functions given by

$$g_2(0)=C$$

$$g_1(n, s)=g_2(n, s) \qquad \text{for } n=1, \ldots, N;\ s=1, 2, \ldots;$$

where $g_1(n, s)$ is nondecreasing in s for each fixed n, Derman and Lieberman show that for a fixed N there exists a sequence of numbers $s_1, s_2, \ldots, s_N$ such that the optimal solution for minimizing th expected average cost-per-unit time is a stationary policy of the form

$$R_{(n, s)}=\begin{cases} 1 \text{ if } s< s_n, \\ \\ 2 \text{ if } s \geqslant s_n. \end{cases}$$

Under additional assumptions on the cost functions, in order to determine the optimal N, only finite number of possible choices for N need to be investigated.

This joint maintenance-inventory model is generalized by Ross [1969a] to allow a continuous sta space and deterioration of the component from period to period. He establishes optimal policies of th same form as Derman and Lieberman and shows that when the inventory is 0, it is optimal to ord N^* units.

2. CONTINUOUS TIME MAINTENANCE MODELS

In this section maintenance models are considered which do not contain the assumption that mai tenance or inspection activity is *a priori* restricted to a particular discrete set of points in time. Th is to say, in these models the actions of the decisionmaker may potentially take place anywhere on th continuous time axis (although there may be only a discrete number of such actions).

In certain of these models, the maintenance activity is permitted to occur as a continuous strea That is, the decisionmaker must optimize over functions $m(\cdot)$ where $m(t)$ is the maintenance expenditu rate at time t. These models are considered first.

ntrol Theory Models

Determination of the optimal maintenance schedule and the sale date for a unit in a deterministic ironment has been under study by a number of different authors. Early solutions to the problem are oe found in Masse [1962]. Naslund [1966] was the first to solve the problem by making use of the ximum principle.

In more recent work, Thompson [1968] presents a simple maintenance model which illustrates the lication of the maximum principle technique to maintenance problems. The model contains the owing factors: the (unknown) sale date of the unit T, the present value $V(T)$ of the unit if its sale date is T, the salvage value $S(t)$ of the unit at time t, the net operating receipts $Q(t)$ at time t, the rate of interest r, the number of dollars spent on maintenance $m(t)$ at time t where maintenance refers to money spent over and above necessary repairs, the maintenance effectiveness function $f(t)$ at time t (in units of dollars added to $S(t)$ per dollars spent on maintenance), the obsolescence function $d(t)$ at time t (dollars subtracted from $S(t)$), and the production rate p at time t.

It is assumed that d, f, and m are piecewise continuous, d is nondecreasing and f is nonincreasing, constant over time, and, for some given constant M, $0 \leqq m(t) \leqq M$.

The object of this model is to choose a maintenance policy $m(t)$ and the sale date T^* to maximize $V(T)$. Thompson begins by solving for the optimal maintenance policy for a fixed sale date T.

He shows that the Hamiltonian is linear in control m, hence the optimal maintenance policy will be of the bang-bang type (piecewise constant). The optimal maintenance policy is obtained by solving $f)=r/(p-(p-r)e^{-r(T-t)})$ for the unique point T'. Then

$$m(t)=\begin{cases} M, & t<T' \\ \text{arbitrary}, & t=T' \\ 0, & t>T'. \end{cases}$$

It is clear that $m(t)$, the optimal control, is a function of T. In this respect, it is piecewise constant; hence, the solution of $\partial V/\partial T=0$ for T is simplified. Thompson presents the details of this procedure, illustrates the model with a few examples and extends the model to the case of a variable production rate.

Arora and Lele [1970] extend Thompson's model by considering the effect of technological progress. This was accomplished by including a term for obsolesence, due to such progress, in the state equation for the salvage value of the unit.

Kamien and Schwartz [1971] consider a maintenance model which represents another extension of Thompson's work. In their model, it is assumed that the value of the unit's output is independent of its age while the probability of failure increases with its age. Furthermore, it is assumed that revenue and scrap value are both independent of age. The probability of failure is influenced by the amount of money spent on maintenance according to the following differential equation:

$$\frac{dF(t)}{dt}=(1-u(t))h(t)(1-F(t)).$$

The function $u(t)\in[0, 1]$ is the level of maintenance at time t; $F(t)$ is the time-to-failure distribution provided the unit is given maintenance according to the schedule $u(\cdot)$; $\hat{F}(t)$ is the corresponding dis-

tribution provided the unit receives no maintenance; and $h(t)=\hat{F}'(t)/(1-\hat{F}(t))$ is the "natu failure rate of the unit. The object of maintenance is to reduce the probability of failure. Expec revenue from the unit's output is maximized by selecting the appropriate control function, and optimal sale date T^* is determined.

Kamien and Schwartz characterize the solution to this problem by deriving necessary co tions for the optimal sale date. In addition, they use the maximum principle to characterize necess conditions for optimal control and in particular show that it is not of the bang-bang type. They also pr sufficiency of their necessary conditions. These results are of interest since the underlying proces this case is controlled by influencing the failure rate and not the salvage value as was the case in Tho son's paper.

In two recent papers by Sethi and Morton [1972] and Sethi [1973], the basic one-unit model is tended to the situation of maintaining a chain of units. In addition, conditions for a changing techno ical environment affecting both production and maintenance requirements of future units is assum In this dynamic model, prices of future units are also allowed to vary. In the first paper, a finite hori problem is considered and a solution procedure for determination of the optimal maintenance sched for each unit in the chain is derived. In addition, conditions for bang-bang control are discussed. In second paper, the problem is solved where the maintenance policy is assumed to be stationary in sense that the same maintenance policy is applied to each unit in the chain. The optimal maintena schedule is characterized and again conditions for bang-bang control are presented. Finally, the co putation of the optimal replacement period is posed in terms of a nonlinear programming problem.

In a related paper by Tapiero [1973], a sequence of n units is also considered. This paper is a dir generalization of Thompson's model. Characterization of optimal maintenance schedules and a cussion of replacement times are presented. Tapiero demonstrates that the decision to replace a u depends only on the relative value of the current unit and the subsequent unit. Thus a replacemen effected when the subsequent unit becomes more profitable to operate. Tapiero refers to this condit as "technical obsolescence."

The above maintenance policies permitted a (piecewise) continuous rate of maintenance experi ture through time. Consider now policies that apply a maintenance action only at discrete instants time; for example, replacing an item upon its reaching a certain age or upon its third breakdown. Su policies can nonetheless be distinguished from those in Part 1 because the event precipitating the ma tenance action is, in general, permitted to occur anywhere on a continuous time axis.

Age Dependent Replacement Models

In the earlier models of age replacement (cf., Barlow and Proschan, 1965) the replacement of unit at failure costs c_2 while replacement before failure costs $c_1 < c_2$. It was shown that if F, the tribution of time-to-failure, had a strictly increasing failure rate then there existed a unique T^* su that expected cost-per-unit time was minimized if the unit was replaced at age T^* or at failure, whi ever occurred first.

Glasser [1967] has obtained solutions to the age replacement problem for three specific distri tions, the truncated normal, the gamma, and the Weibull.

Fox [1966b] showed the optimality of an age replacement policy under a total discounted c criterion. For a continuous and strictly increasing hazard rate, he derives an integral equation wh can be solved for the optimal T^*.

Schaefer [1971] extends the standard age replacement model by including an age-dependent cost. ich a cost may reflect the increasing burden of routine maintenance as the unit ages, its diminishing oductivity, or reduced salvage value for the unit at replacement due to depreciation. Specifically, expresses the total cost up to time t as

$$C(t) = c_1 N_1(t) + c_2 N_2(t) + c_3 \left[\sum_{i=1}^{N(t)} Z_i^{\alpha} + (t - S_{N(t)})^{\alpha} \right],$$

ere $N_1(t)$ is the number of replacements due to reaching the age replacement level T which have curred by time t, $N_2(t)$ is the number of replacements due to failure by time t, $N(t) = N_1(t) + N_2(t)$, > 0, Z_i is min (X_i, T) where X_i would be the uninhibited life of the ith unit in the sequence of re- acements and T is the fixed age at which replacement is to be made, S_k is the time of kth replacement, d $0 < \alpha < 1$. The goal is to minimize the long run average cost-per-unit time. By an argument similar the one given in Chapter IV of Barlow and Proschan [1965], it can be shown that the optimum policy nonrandom if the failure distribution F is continuous. The case of an exponential life distribution analyzed in more detail.

R. Cleroux and M. Hanscom [1974] considered a very similar model. One of the differences is that e age-dependent cost $c_3(ik)$ is incurred only at discrete times, the multiples of the positive constant Moreover, $c_3(ik)$ need not be increasing in i. For F continuous and IFR, they show that the optimal e replacement policy is nonrandom for an infinite time span. They then develop sufficient conditions the optimal replacement interval T^* to be a finite number and to be restricted to a certain finite set.

The notion of an age-dependent cost structure is further generalized by M. R. Wolfe and R. Subra- nian [1974]. If the nth unit is replaced at age T the total cost incurred over the life of that unit is

$$W_n = \int_0^T [Y_n + r(s)]ds + K_n,$$

ere Y_n and K_n are independent random variables forming renewal processes and $r(\cdot)$ is a differentia- and strictly increasing function. K_n is the cost of replacement and $Y_n + r(s)$ is the cost rate at time fter the installation of the nth unit. There are no failures. In order to minimize the expected total t-per-unit time, the decisionmaker determines a critical threshold value c^* such that when the cost e exceeds c^* he replaces the unit. For a particular realization of Y_n, this occurs at age $r^{-1}(c^* - Y_n)$. cedures for determining c^* are given, and for $r(t)$ linear, they derive an explicit solution.

An example where an optimum age replacement policy is found for a two-unit redundant system rovided by T. Nakagawa and S. Osaki [1974]. While one of the identical units is in operation, the er is in standby status, immune to all failure or aging effects. When the operating unit is sent to re- r, either for preventive maintenance (which renews the unit) or because of failure, the standby unit es over. If an operating unit should fail when the other unit is still at the repair facility, the system omes inoperative and the most recent failure must wait its turn for repair. Under the assumption that reventive maintenance activity for a unit entails less mean time in repair than does the repair of a 's failure, it may be advisable to routinely schedule such preventive maintenance after a unit has n in operation for time T. (If at that point in time the operating unit lacks a standby, preventive main- ance is delayed until the current repair work is completed.) Nakagawa and Osaki derive the optimal

value for this T under the assumptions of increasing failure rate and with regard to maximizing the long run proportion of time the system is operating. Their arguments utilize the regenerative propert of the system at those epochs when the operating unit enters the repair facility and a standby unit is available to take over.

With very expensive and complicated systems, the failure of a single component unit would very reasonably call for replacing the entire system. Instead the system could be restored to operati by replacing the single failed unit. Because the great bulk of that system's components were not newed, the probability distribution for the system's remaining life remains essentially what it was at instant before failure. That is, the failure and subsequent repair activity often do not affect the syste failure rate. The action of restoring a failed system to operation without affecting its failure rate is cal minimal repair. Barlow and Hunter [1960] incorporated this notion in their Policy Type II replacem Model (Policy Type I was the simple age replacement model—replace at age T or failure, whiche comes first). A Type II Policy assumes that the unit is replaced after functioning for T units (downti is not included in the T). Any failures before that time would be dealt with by minimal repair. The o mization of a Type II policy with instantaneous minimal repair is equivalent to the age replacem model which incorporated an age-dependent cost of operating the unit. This age-dependent cost is expected cost of incurring the expense of minimal repair which depends on the unit's age via the fail rate. Bellman [1955] and Descamp [1965] applied dynamic programming to this problem. Sivazl [1973] generalized their work by permitting a positive downtime for the minimal repair followin failure. This downtime is random with an arbitrary distribution. Using the functional equation techni he derives an explicit expression for the long term total expected cost. Further, he derives necessary a sufficient conditions for the optimal policy to be of the "Type II" form described above.

Makabe and Morimura [1963a, 1963b, 1965] and Morimura [1970] introduce a Policy of Type and Type III′. Under Policy Type III, the unit is replaced at the kth failure. The $k-1$ previous failu are corrected with minimal repair. A Policy Type III, Morimura feels, would be easier to implement many practical situations than a Type II policy.

An optimal policy of Type III is shown to exist for strictly IFR distributions, both with respect the criterion of expected fraction of time operating in $[0, \infty)$, called limiting efficiency: and with resp to what Makabe and Morimura [1963a, 1963b] call the maintenance cost rate which is defined to b

$$\begin{aligned}&[\text{cost-per-unit downtime}] \times [\text{expected fraction of downtime}]\\ &+ [\text{expected cost of all repairs and replacements during a unit time}].\end{aligned}$$

Also for distributions of strictly increasing failure rate, Morimura [1970] has shown the existe of an optimal policy (with respect to limiting efficiency) for a larger class of policies, Type III′ polici This class of policies is specified by two critical numbers t^* and k. All failures before the kth failu are corrected only with minimal repair. If the kth failure occurs before an accumulated operating ti of t^*, it is corrected by minimal repair and the next failure induces replacement. But if the kth failu occurs after t^*, it induces replacement of the unit. Clearly, if $t^* = 0$, this reduces to a policy of Type I.

Even if the replacement rule is purely according to age, it cannot always be assumed that the p vision of replacement units is outside the purview of the decisionmaker. Falkner [1968] examines maintenance-inventory problem where there is a single unit which is operating and it fails according an IFR distribution $F(\cdot)$. When the unit reaches a certain age or when it fails, it is replaced by anoth

w unit. The problem is to find both the initial number of spares, N, to produce and the age replacement policies $\tau_j(t)$, $j=1, \ldots, N+1$, for the original new unit and the spares in order to minimize the expected total operating cost of the system over a finite time interval $[0, T]$. This problem is formulated as a dynamic program with nondecreasing costs for holding stock (h per unit), stockout penalty (p per unit) and unscheduled replacement (r per unit). Under the assumption $F(0) = 0$ Falkner shows that the optimal number of spares is bounded above by the greatest integer less than or equal to p/h, where $h > 0$. With additional assumptions on $F(\cdot)$ he is able to characterize the cost structure and the structure of the optimal age replacement policies.

In a later paper, Falkner [1969] specializes this model by removing the age replacement decision process and is able to obtain stronger results on the optimal initial number of spares. He gives an application using the negative exponential failure distribution.

Shock Models

In most maintenance models the time-to-failure random variable of a unit is considered intrinsic to that unit. But it is possible to take a different view—regarding the unit as being subject to exterior shocks, each of which damages (or causes wear) in such a way that the damage accumulated up to a particular time defines the unit's probability of failure at that time.

For example, A-Hameed and Proschan [1973] set up the shock process as a nonhomogeneous Poisson process. This process, combined with the probabilities P_k that the unit will survive k shocks, induces a time-to-failure distribution for the unit. Various properties of the distribution $\{P_k\}$ are related to corresponding properties in the induced time-to-failure distribution. For example, if the former distribution is IFR, so is the latter.

The question of optimal replacement rules in the context of a shock model has received attention only very recently. Taylor [1973] considers a unit subject to a shock process where the decisionmaker knows at all times the level of accumulated damage from the shocks. At the occurrence of a shock he has the option of replacing the item at cost c_1. If the item should fail, it must be replaced at cost $c_2 > c_1$. The shocks occur according to a Poisson process and each shock causes a random amount of damage which accumulates additively. The device may fail only at the occurrence of a shock and then with a probability which depends on the accumulated damage. If the probability of the device failing is an increasing function of the accumulated damage, Taylor proves that the optimal replacement rule is of the following control limit form: replace the device at failure or when accumulated damage first exceeds a critical control level Z^*. He gives an equation which implicitly defines Z^* in terms of the replacement costs and other system parameters.

Richard Feldman [1974] and [1975] has permitted a more general stochastic process to represent the incidence of damage to the unit. In particular, both the time to the next shock and the degree of damage inflicted by that shock may depend on the current level of accumulated damage. He then derives the optimal control limit rule for replacement.

Interacting Repair Activities Models

For a system composed of many units, the repair or replacement of one unit should sometimes be considered in conjunction with what happens to the other units. Below are discussed four ways in which

a maintenance policy either gives rise to, or exploits, interactions among the units of a system; name opportunistic policies, cannibalization policies, multistage replacement policies, and variable repair r policies.

Opportunistic policies exploit economies of scale in the repair or replacement activity. That two or more repair activities done concurrently may cost less than if they are done separately. So t necessity of performing at least one repair might provide the economic justification to do several oth at the same time.

In cannibalization and multistage replacement models, units of the same type are utilized at c ferent locations in the system. In response to a failure at one location, an identical unit may be tra ferred there from another location. In the multistage model, a new item must enter the system at so location so that all units are restored to operation. Also in multistage models, the purpose of tra ferring units among locations is to locate those units, which because of their age are less likely to fa at those locations where failure is most costly. In the cannibalization model, on the other hand, no n item enters the system when a transfer is made. Since the performance of the system depends on whi items are functioning, the purpose of the transfer is to provide the system with the best possible c figuration of functioning units.

The last type of interacting activities model to be discussed is a variable repair rate model; that when the repair capacity of a system is limited and under the decisionmaker's control, he may wish modify that capacity according to the number of items in a down state.

Some recent work has examined opportunistic repair in the context of the following system str ture: There are two classes of components, 1 and 2. Class 1 contains M standby redundant compone so that upon the failure of the currently operating class-1 component, a standby takes over. When the class-1 standbys have failed, the system suffers catastrophic failure. The class-2 components, the other hand, form a series system; if one of them should fail, the system suffers a minor breakdo The operator always knows the state of the system.

When a minor breakdown occurs, there is the opportunity for opportunistic repair of those clas items which have failed. Kulshrestha [1968a] examined such a policy under the assumption that t class-1 units fail according to a general distribution and the ith component of class-2 fails with the c stant rate λ_i ($i=1, 2, \ldots, N$). The time to complete repairs follows a general distribution. He th compares this policy to the corresponding nonopportunistic policy upon assuming the failure distri tion of the class-1 units to be exponential.

Nakamichi, et al. [1974] examine the same system, but they require that any class-1 unit enter pair upon its own failure—not waiting for either a minor or major system failure. They further assu that in a major breakdown, as soon as the class-1 unit being serviced is repaired, it is reinstalled p mitting the system to operate again. And in a minor breakdown, as soon as the failed class-2 unit is paired the system operates again. In the case where a class-2 unit fails when a class-1 unit is under pair, two alternatives are examined. (1) It is repaired immediately, interrupting the repair of the clas unit and (2) it's repair awaits the completion of the class-1 unit's repair.

The notion of cannibalization stems from the situation where a working (operative) part in a syst may be removed from one location to replace a failed part in another location. This "cannibalizatio would be done in order to improve the operation of the system in some sense.

The foundations for the study of cannibalization were established by Hirsch, Meisner, and B [1968]. Each part in the system is classified by type ($1, \ldots, N$) and location ($1, \ldots, n$). The set

ts is then partitioned into subsets such that any two parts in a subset are interchangeable, but parts n different subsets are not. At any point in time, the status of all parts in their locations is given by a ıry n-tuple $v=(v_1, \ldots, v_n)$,

re

$$v_i = \begin{cases} 0 & \text{implies part } i \text{ has failed} \\ 1 & \text{implies part } i \text{ is operating.} \end{cases}$$

assumed that failures are detected immediately. The state of the system is given by a monotonic easing "structure" function $\varphi(v)$ which takes values from $\{0, 1, \ldots, M\}$, where 0 is total system re and M is the best performance. In the event there are spare parts available and a failure occurs, failed part is immediately replaced by a spare part and the structure function is unchanged. How-, if there are no spares, then part shortages exist and the problem is to find cannibalizations (trans- ıations that make feasible interchanges of parts) which maximize φ over all feasible interchanges. cannibalization T which maximizes φ is called admissible. Under an assumption, known as the nimum condition," on the composite transformation $\varphi \cdot T \equiv \varphi T$, Hirsch, Meisner, and Boll ex- tly characterize the state of the system for any admissible cannibalization T as a function of the ber of working parts of each type. The minimum condition asserts that $\varphi T(v)$ is equal to the mini- ın value of $\varphi T \pi_i(v)$ over $i = 1, \ldots, N$, where $\pi_i(v)$ is the operation of making all parts of type j able (for all $j \neq i$) while the status of parts of type i are held constant. Thus, in a sense, a single type determines the value of φT for any v. Using the minimum condition and the explicit charac- ation of the state of the system, φT, for any admissible T, they demonstrate the probability laws ıe state of the system under the additional assumptions that (i) the failure distributions of parts pe i are identically distributed and do not depend on the location of the part for each $i = 1, \ldots,$ ıd (ii) the lifetimes of all parts are independent random variables.

Simon [1970, 1972] generalizes the results of Hirsch, Meisner, and Boll by relaxing the restric- on the interchangeability of parts. He classifies their interchangeability as closed, isolated, or communicating classes or parts. When the closure, isolation, or communication aspects are ked, then the result of Hirsch, Meisner, and Boll that all admissible cannibalizations are equivalent nger holds. The question becomes: from the set of admissible policies for a given v, are there poli- which are "better" than others? With the additional objective of maintaining the most flexibility uture cannibalization, Simon [1970] demonstrates that certain admissible policies are uniformly r than others. In another paper [1972] Simon establishes upper and lower bounds on φT under ıore general interchangeability rules and from these bounds he also develops bounds on the proba- laws of the state of the system.

In a later work, Hochberg [1973] returns to the interchangeability classes of Hirsch, Meisner, Boll; however, rather than allowing each part at its location to occupy only one of two states—failed erating (i.e., 0-1), he assumes it can be in any one of k states

$$\{a_i\}_{i=1}^{k} \text{ where } 0 = a_k < \ldots < a_2 < a_1 = 1.$$

successively smaller state represents a decreasing level of performance. With this generalized iption of the status of each part at any point in time and the minimum condition, Hochberg (paral-

leling the work of Hirsch, et al.) obtains a characterization of the state of the system φT (i.e., the p formance level over admissible cannibalizations) as a function of the number of working parts at ea level a_k. He then develops the probability distribution of φT.

Implicit in the Hirsch, Meisner, and Boll, Simon, and Hochberg works is the assumption that parts start operating from time zero and are continuously subject to failure as long as the system op ates. For many systems when a working part is in a major assembly that has failed for other reaso this working part does not experience any further deterioration or stress until it is cannibalized a commences to operate again. Rolfe [1970] looked at this aspect of cannibalization. He considered group of S identical major assemblies which operate independently of each other. A major assemb contains N distinct subassemblies (or parts). Each part is interchangeable with its corresponding pa in the other major assemblies, but with no other parts. All working assemblies operate continuously f a period of time T before they are inspected. During this period of operation, parts may fail but the maj assemblies continue to function. Upon inspection, failed parts are immediately replaced from an initi stock of spares until the stock is exhausted after which they are replaced by cannibalized parts fro other nonworking major assemblies. Assemblies containing any failed parts are not used in the ne period of operation. It is assumed that all parts fail independently while operating and that parts of ty i have failure distribution $1-e^{\lambda_i t}$. With these assumptions the state vector of the stochastic process described by the number of working parts of each type $n_i(t)$ available at the end of each operating pe iod t. This stochastic process forms an S^N+1 state, absorbing Markov chain where the absorbing sta is reached when $n_i(t)=0$ for some $i=1, 2, \ldots, N$. Rolfe develops the expected number of good maj assemblies at the end of any period and because of obvious computational difficulties when S and/c N are large, he gives approximations and lower bounds for this expectation.

As mentioned before, multistage models differ from cannibalization models in that new items ent the system to replace items transferred to other locations. Bartholomew [1963] examined the followi model: Suppose a system contains N units with independent but identically distributed times-to-failur These N units, although stochastically identical, are partitioned into two classes, class I and class (according, perhaps, to their function or location within the system). Upon the failure of any unit, must be replaced—at a cost k_1 for items in the first class and at cost k_2 for items in the second class.

The procedure Bartholomew analyzes, the so-called two-stage replacement strategy, is to repla all failures that occur amongst items in class II with items in class I and to replace items in cla I, which either failed or were transferred to class II, with new items. There is a cost β for transferri one item from class I to class II. Also, there is a purchase cost per item (which turns out to be relevant). It is assumed that replacing and transferring items takes no time.

The total number of unit failures remains unaffected by the above procedure. But the proporti of failures in class I vis-a-vis class II may change. Bartholomew determines the following conditi under which a two-stage replacement strategy is better than simple replacement at failure.

If ρ is the transfer rate from class I to class II, n_i is the number of units of class i, and μ is the me time-to-failure of a unit; then the two-stage strategy is preferable if

$$n_2(k_2-k_1)\mu^{-1}-n_1\rho(k_2-k_1+\beta)>0.$$

The determination of ρ depends on whether an item from class I is selected for transfer to class II on random basis or by an oldest-first rule. Upon assuming $\beta=0$, he derives an approximation for ρ for ea of these cases.

One implication of his results is that for IFR time-to-failure distributions, $k_1 > k_2$ makes the two-age scheme preferable to simple replacement at failure. In the DFR case with $k_1 < k_2$, the two-stage scheme is also better.

Naik and Nair [1965a, 1965b] have generalized the above scheme to multistage replacement strategies. Marathe and Nair [1966b] investigate multistage block replacement strategies. This latter model requires the assumption that there is a reserve of units, the "interstage inventory," attached to each class.

In his model of a variable repair rate problem, Crabill [1974] considers a collection of $M+R$ units. The state of the system, $i(i=1, \ldots, M+R)$, designates the number of these units in operating condition. A cost-per-unit time of $C(i)$ is charged when in state i. Min(i, M) of the i units in operating condition are actually in production, and only these are subject to failure—at the constant hazard rate λ. When a unit fails, it enters the single-server repair facility (or its queue). On the basis of the current state of the system, the decisionmaker selects action $k(k=1, \ldots, K)$ which provides the repair facility with an exponential repair rate u_k at per-unit-time cost n_k. The object is to minimize long run cost-per-unit time. Using Markov decision theory, Crabill presents sufficient conditions for particular service rates to be eliminated from consideration regardless of state. Furthermore, he provides sufficient conditions ensuring that the optimal service rate is a nonincreasing function of the system's state.

Incomplete Information

As with discrete time models, the decisionmaker may lack complete information about any of the following: the current state of the system (unless he performs an inspection), the probability law governing the system's stochastic behavior (e.g., the time-to-failure distribution of a single unit), and the cost implications of particular operating policies.

When the current state of the system is not known, the problem arises of jointly determining a replacement and inspection policy. In a model by Savage [1962] the state of the unit moves from $x=0$ (new unit) to $x=k$, $k=1, 2, \ldots$, according to a Poisson process. In state x, income is earned at the rate $i(x)$. At a cost L, the decisionmaker can inspect the unit and thereby learn its true state. After each inspection, he elects either to schedule another inspection $T(x)$ time units into the future and not replace the unit at present, or to replace the unit (thus returning to $x=0$), with the next inspection $T(0)$ time units into the future. Replacement requires m units of time at a cost of c per unit.

The objective of the decisionmaker is to minimize the long run average cost-per-unit time by specifying the "how-long-to-next-inspection" function $T(x)$ and the set W of all states which will call for another inspection rather than replacement. Savage shows that if $i(x)$ is nonincreasing, then $T(x)$ is strictly decreasing for x in W and bounds on $T(x)$ are derived. More explicit results are derived for two special cases of $i(x)$.

Antelman and Savage [1965] consider a parallel problem, namely, the process governing the change of states is Brownian Motion rather than a Poisson process. In particular, it is then possible to move to an improved state without a replacement.

These models were generalized by Chitgopekar [1974]. He considers a larger (but still finite) action space which includes the class of all random time-to-inspection policies while permitting a more general stochastic process to govern the change of states. He shows that the optimal policy is nonrandom.

Keller [1974] utilized optimal control theory for an approximate method of selecting that inspectio schedule, for a unit subject to failure, which will minimize cost until the detection of the first failure Each inspection has cost L and the cost $H(t)$ is incurred when detection of the failure by an inspectio occurs time t after failure. This problem is placed in the control theory framework by assuming tha the tests are so frequent that they can be described by a smooth density $n(t)$ which denotes the numbe of checks per unit time. In other words, at time t, the tests are scheduled $1/n(t)$ units of time apart He then derives an integral equation for $n(t)$ and uses this solution to minimize the expected cos up to detection of the first failure.

Several authors consider the problem of incomplete information concerning the time-to-failur distribution governing the system. A single unit is subject to random failure which can be detected witl probability p by an inspection. Each inspection has cost L and $v \cdot t$ is the cost of a failure which remains undetected for a duration t. The problem is to schedule inspections so as to minimize these two cost over the period the unit remains installed which can be no longer than the time when failure is firs discovered by an inspection.

Derman [1961] has found a minimax optimal solution to this problem when the time-to-failure distri bution is totally unknown. However, it was necessary to assume a finite time horizon T; i.e., the cos accounting stops at either the first inspection to detect failure or at time T, whichever happens first The reason for this is that for any possible inspection schedule there exists a distribution which woul induce an arbitrarily high expected cost during an infinite time horizon. Hence, a minimax solutio would not then exist. McCall [1965] gives Derman's results.

Roeloffs [1963, 1967] (letting Derman's probability of detection, p, equal one), found the minima inspection schedule when a single percentile is all that is known of the time-to-failure distribution That is, y and π are known where $F(y) = \pi$. He finds the minimax schedule $(x_1, x_2, \ldots, x_{m+n})$ where $0 \leq x_1 \leq \ldots \leq x_m \leq y \leq x_{m+1} \leq \ldots \leq x_{m+n} \leq T$, and the corresponding expected cost for a cos structure identical to Derman's. Naturally, the expected costs are less than Derman's due to the addi tional information. It is interesting to note that the form of his solution for selecting the inspection point after y, is identical to that of Derman. That is to say, the information contained in the percentile has no bearing in carrying out a minimax surveillance after that percentile is crossed. Roeloffs also finds the optimal inspection schedule to minimize expected cost-per-unit time (rather than per period of installation). However, in doing this, he sets $T = y$. Kander and Raviv [1974] utilize dynamic pro gramming to model this problem for an arbitrary T.

Combining the unknown time-to-failure distribution with the traditional age replacement problem (cost c_2 for replacement at failure, cost $c_1 < c_2$ for scheduled replacement), Fox [1967] used a Bayesian approach. As the realizations of times-to-failure or scheduled replacements are progressively observed the decisionmaker learns more about the underlying time-to-failure distribution. Using the Weibull time-to-failure distribution with the gamma *a priori* distribution for the parameter, his objective is to minimize the long run expected discounted cost. He derives certain asymptotic optimality conditions on the stationary policy for an arbitrary fixed number of replacements.

In actual maintainability applications, the exact form of the function relating expected cost to the control variable is often not known. In such cases it may be appropriate to precede any attempt a optimization by an experiment which "samples" the resultant costs for various values of the contro variable. Then, using least-square techniques, estimate the shape of the actual cost function—thereby permitting an inference as to the optimal value for the control variable.

Such a model was proposed by Dean and Marks [1965] and analyzed by Elandt-Johnson [1967]. routine maintenance is provided a machine (or vehicle fleet) with frequency x per year, the average sulting cost is $C(x) = bx + D(x)$ where b is the average cost of a scheduled maintenance and $D(x)$ the unknown expected cost-per-unit time of providing emergency nonscheduled maintenance. Pre- mably, $D(x)$ is decreasing. $C(x)$ is assumed to possess an absolute minimum, say at x_0. For the ecific form

$$D(x) = E\left[\sum_{k=0}^{n} A_k x^k + \Theta\right],$$

ere Θ is a random variable with distribution $N(0, \sigma^2)$ and the A_k's are not known, least-square timates (a_k) are obtained for the coefficients (A_k). By those estimated coefficients, the estimated al cost function can be minimized for $\hat{x}_0$.

Elandt-Johnson [1967] provides a normal distribution which approximates the true distribution the statistic $\hat{x}_0$. She further shows that $E[C(x_0) - C(\hat{x}_0)]$ depends on the degree and coefficients of e polynomial

$$\sum_{k=1}^{n} A_k x^k.$$

REMARKS

The foregoing survey has been primarily expository rather than critical. This approach was taken order to be able to include most available works on maintainability in a reasonably sized monograph. was felt that the more papers included, the more useful it would be to a general reader, practitioner, research scholar or teacher interested in entering the area.

Now that the reader has seen what has been accomplished in maintainability, the question be- mes, "what new work needs to be undertaken?"

In the case of deteriorating single-unit inspection, repair and/or replacement models, there have en many generalizations of the basic Derman models. For example, the state space has been allowed become nondiscrete, the costs a nonstationary random variable and the sojourn time in a state a dom variable. It would appear that most future research on these types of models would be to add re system constraints and to develop efficient algorithms to solve the large linear, nonlinear or amic programs involved, rather than to provide additional refinements to the assumptions of the sic model. Such efforts will tend to be directed toward particular problems in the sense that the dels will become more specialized and tailored to particular situations. Similar comments hold for th area of age-dependent replacement models and the maintenance-salvage models.

The area requiring the most future work is the study of multiechelon multipart interaction mainte- ice models. This area is perhaps the most difficult to handle mathematically especially when the ir ractions occur because of stochastic dependence among the parts. In many such cases, very little c be said about the operating characteristics or optimal solutions of the models. Often the only way th se problems can be handled is by simulation; consequently, few general results will be obtained in th near future. In the long run, however, work will be initiated to handle more complex kinds of sto- cl stic dependence, where the dependence is restrictively proscribed.

Multipart, multiechelon models which involve economic dependencies are often less difficult t model and solve than the stochastic dependencies. Frequently, as in the inventory-maintenanc models, large mathematical programs result and the question becomes how to solve them rather tha how to model or describe the problem. Future research will see an increase in the general applic bility of these models. Of course, making them more general and/or realistic usually removes the opt mality of simple, elegant policies and one is dependent on large computers to solve the complicate mathematical programs involved. Fortunately, many decisionmakers in the military branches, othe governmental agencies, industry, and private nonprofit organizations now possess the sophisticatio to use more complicated techniques and approaches. The phenomenal progress of space age technolog is to be thanked for this development, as well as the need to maintain more increasingly comple devices.

Perhaps the most future research will go into the development of new models with different cor straints that are needed to handle maintenance problems in health, ecology, and the environmen With the advent of a national health insurance program, there will be an increased effort nationwid on preventive medicine in order to cut costs and to extend the lifetimes of previously disadvantage groups. Similarly, with the depletion of many of our plant and animal resources, new concepts o maintainability must be developed to restore balances for future populations. Finally, there are man environmental problems that require inspection and corrective action. Models for all of these problem will be built. Some general theory for these models will be established and large scale versions of th problems will be solved.

ACKNOWLEDGMENT

We thank Professor Morris Cohen of the Wharton School, The University of Pennsylvania, fo his contribution to the initial write-up of the optimal control of continuous time maintenance model with salvage value, and Professor Evan Porteus of the Graduate School of Business, Stanford Uni versity, for his excellent review of this work.

BIBLIOGRAPHY

[1] A-Hameed, M. S. and F. Proschan, "Nonstationary Shock Models," Stoch. Proc. Appl. *1*, 383–40 (1973).

[2] Allen, S. G. and D. A. D'Esopo, "An Ordering Policy for Repairable Stock Items," Oper. Re *16*, 669–675 (1968).

[3] Aroian, L. A., T. I. Goss, and J. Schmee, "Maintainability Demonstration Test for the Parameter of a Lognormal Distribution," AES–747, Institute of Admin. and Mgmt., Union College, Sche nectady, N.Y. (1974).

[4] Arora, S. R. and P. T. Lele, "A Note on Optimal Maintenance Policy and Sale Date of a Machine," Man. Sci. *17*, 170–173 (1970).

[5] Arrow, K., D. Levhari, and E. Sheshinski, "A Production Function for the Repairmen Problem," The Review of Economic Studies *39*, 241–249 (1972).

[6] Arunkumar, S., "Nonparametric Age Replacement Policy," Sankhya, Series A, *34*, Pt. 3, 251–25 (1972).

7] Avramchenko, R. F., "Optimum Scheduling of the Use of Spare Elements in the Loaded Mode," Eng. Cybernetics *8*, 480–483 (1970).

8] Bakut, P. A. and Yu. V. Zhulina, "A Two-Stage Procedure of Decision Making," Automatika i Telemekhanika *8*, 156–160 (1971).

9] Bansard, J. P., J. Descamps, G. Maarek, and G. Morihain, "Stochastic Method of Replacing Components of Items of Equipment which are Subject to Random Breakdowns: The 'Trigger' Policy," Metra *10*, 627–651 (1971).

0] Barlow, R. E. and L. Hunter, "Optimal Preventive Maintenance Policies," Oper. Res. *8*, 90–100 (1960).

[1] Barlow, R. E., F. Proschan, and L. C. Hunter, *Mathematical Theory of Reliability* (New York, Wiley, 1965).

2] Barlow, R. E. and P. Chatterjee, "Introduction to Fault Tree Analysis," Oper. Res. Center, College of Engineering, University of California at Berkeley, ORC 73–30 (1973).

] Barlow, R. E. and F. Proschan, "Importance of System Components and Fault Tree Events," Oper. Res. Center, College of Engineering, University of California at Berkeley, ORC 74–3 (1974).

] Bar-Shalom, Y., R. E. Larson, and M. Grossberg, "Application to Stochastic Control Theory to Resource Allocation under Uncertainty," IEEE Trans. Aut. Control *AC–19*, 1–7 (1974).

] Bartholomew, D. J., "Two-Stage Replacement Strategies," Oper. Res. Quart. *14*, 71–87 (1963).

] Barzilovich, Ye. Yu., V. A. Kashtanov, and I. N. Kovalenko, "On Minimax Criteria in Reliability Problems," Eng. Cybernetics *9*, 467–477 (1971).

1] Beja, A., "Probability Bounds in Replacement Policies for Markov Systems," Man. Sci. *16*, 257–264 (1969).

1 Bellman, R., "Equipment Replacement Policy," SIAM J. on Appl. Math. *3*, 133–136 (1955).

1 Bellman, R., *Dynamic Programming* (Princeton University Press, Princeton, N.J., 1957).

2 Bellman, R. and S. Dreyfus, *Applied Dynamic Programming* (Princeton University Press, Princeton, N.J., 1962).

2 Bhattacharyya, M. N., "Opt. Allocation of Stand-By Systems," Oper. Res. *17*, 337–343 (1969).

2 Blanchard, B. S., "Cost Effectiveness, System Effectiveness, Integrated Logistics Support, and Maintainability," IEEE Transactions on Reliability *R–16*, 117–126 (1967).

2 Blumstein, A., R. Cassidy, W. Gorr, and A. Walters, "Optimal Specification of Air Pollution Emission Regulations including Reliability Requirements," Oper. Res. *20*, 752–763 (1972).

2 Bracken, J. and K. Simmon, "Minimizing Reductions in Readiness Caused by Time-Phased Decreases in Aircraft Overhaul and Repair Activities," Nav. Res. Log. Quart. *12*, 159–165 (1965).

2 Brandt, E. and D. Limaye, "MAD: Mathematical Analysis of Downtime," Nav. Res. Log. Quart. *17*, 525–534 (1970).

2 Brown, D. B. and H. F. Martz, Jr., "A Two-Phase Algorithm for the Maintenance of a Deteriorating Component System," AIIE Transactions *2*, 106–111 (1970).

27 Brown, D. B. and H. F. Martz, Jr., "Simulation Model for the Maintenance of a Deteriorating Component System," IEEE Transactions on Reliability R–*20*, 28–32 (1971).

28 Bryson, A. E. and Y. C. Ho, *Applied Optimal Control* (Blaisdell, Waltham, Mass., 1969).

29 Chatterjee, P., "Fault Tree Analysis: Min Cut Set Algorithms," Oper. Res. Center, College of Engineering, University of California at Berkeley, ORC 74–2 (1974).

[30] Cherenkov, A. P. and P. I. Popov, "Fluxes of Failures of Technological Systems Taking Accou of Prophilaxis," Automatika i Telemekhanika *6*, 144–148 (1970).

[31] Cherenkov, A. P. and A. G. Ivanov, "An Iterational Method for Calculating the MTBF of Markovian System," Automatika i Telemekhanika *8*, 143–151 (1972).

[32] Cherniavsky, E. A., "Some Contributions to Failure Models: The Shock and Continuous Dang Processes," Ph. D. Dissertation, Cornell University, Ithaca, N.Y. (1973).

[33] Chitgopekar, S. S., "A Note on the Costly Surveillance of a Stochastic System," Nav. Re Log. Quart. *21*, 365–371 (1974).

[34] Chu, W. W., "Adaptive Diagnosis of Faulty Systems," Oper. Res. *16*, 915–927 (1968).

[35] Cleroux, R. and M. Hanscom, "Age Replacement with Adjustment and Depreciation Costs a Interest Charges," Technometrics *16*, 235–239 (1974).

[36] Cozzolino, J., "Decreasing Failure Rate Processes," Nav. Res. Log. Quart. *15*, 361–374 (1968)

[37] Crabill, Thomas B., "Optimal Control of a Service Facility with Variable Exponential Servi Times and Constant Arrival Rate," Man. Sci. *18*, 560–566 (1972).

[38] Crabill, Thomas B., "Optimal Control of a Maintenance System with Variable Service Rates Oper. Res. *22*, 736–745 (1974).

[39] Cruon, R., A. Rougerie, and C. Van de Casteele, "Scheduling Repairs and Spare Items," *Re Frse. d'Informatique et de Recherche Operationnelle*, Vol. *3*, No. V-3, 87–102 (1969).

[40] Demmy, W. S., "Computing Multi-Item Procurement and Repair Schedules in Reparable-Ite Inventory Systems," Working Paper, Ernst and Ernst, 2300 Winters Bank Building, Dayto Ohio (1974).

[41] Denardo, E. and B. Fox, "Nonoptimality of Planned Replacement in Intervals of Decreasin Failure Rate," Oper. Res. *15*, 358–359 (1967).

[42] Denby, D. C., "Minimum Downtime as a Function of Reliability and Priority Assignments Component Repair," J. Ind. Engr. *18*, 436–439 (1967).

[43] Denicoff, M., S. E. Haber, and T. C. Varley, "Military Essentiality of Naval Aviation Spare Parts Man. Sci. *13*, B439–B453 (1967).

[44] Derman, C., "On Minimax Surveillance Schedules," Nav. Res. Log. Quart. *8*, 415–419 (1961).

[45] Derman, C., "On Sequential Decisions and Markov Chains," Man. Sci. *9*, 16–24 (1962).

[46] Derman, C., "Optimal Replacement and Maintenance under Markovian Deterioration wi Probability Bounds on Failure," Man. Sci. *9*, 478–481 (1963a).

[47] Derman, C., "On Optimal Replacement Rules when Changes of State are Markovian," Chap. *Mathematical Optimization Techniques*, R. Bellman (ed.), (University of California Pre Berkeley and Los Angeles, 1963b), pp. 201–210.

[48] Derman, C. and M. Klein, "Surveillance of Multi-Component Systems: A Stochastic Traveli Salesman Problem," Nav. Res. Log. Quart. *13*, 103–111 (1966).

[49] Derman, C. and G. J. Lieberman, "A Markovian Decision Model for a Joint Replacement a Stocking Problem, Man. Sci. *13*, 609–617 (1967).

[50] Derman, C., *Finite State Markovian Decision Processes* (Academic Press, New York, 1970).

[51] Derman, C., G. J. Lieberman, and S. Ross, "Assembly of Systems Having Maximum Reliability Nav. Res. Log. Quart. *21*, 1–12 (1974).

[52] Descamps, R. and G. Maarek, "Maintenance and Parts Replacement," Gestion *5*, 367–373 (196

3] De Veroli, J. D., "Optimal Continuous Policies for Repair and Replacement," Oper. Res. Quart. *25*, 89–98 (1974).

4] Dishon, M. and G. Weiss, "Burn-In Programs for Repairable Systems," IEEE Transactions on Reliability *R–22*, 265–267 (1973).

5] Downton, F., "The Reliability of Multiplex Systems with Repair," J. Roy. Statist. Soc., B, *28*, 459–476 (1966).

6] Drinkwater, R. W. and N. A. J. Hastings, "An Economic Replacement Model," Oper. Res. Quart. *18*, 121–138 (1967).

7] Eckles, J. E., *Optimal Replacement of Stochastically Failing Systems* (Institute in Engineering-Economic Systems, Stanford University, 1966).

8] Eckles, J. E., "Optimum Maintenance with Incomplete Information," Oper. Res. *16*, 1058–1067 (1968).

9] Elandt-Johnson, R. C., "Optimal Policy in a Maintenance Cost Problem," Oper. Res. *15*, 813–819 (1967).

10] Eppen, G. D., "A Dynamic Analysis of a Class of Deteriorating Systems," Man. Sci. *12*, 223–240 (1965).

11] Esary, J. D., A. W. Marshall, and F. Proschan, "Shock Models and Wear Processes," Ann. of Probability *1*, 627–649 (1973).

12] Everett, H., III, "Generalized Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources," Oper. Res. *11*, 399–417 (1963).

13] Falkner, C. H., "Jointly Optimal Inventory and Maintenance Policies for Stochastically Failing Equipment," Oper. Res. *16*, 587–601 (1968).

14] Falkner, C. H., "Optimal Spares for Stochastically Failing Equipment," Nav. Res. Log. Quart. *16*, 287–295 (1969).

15] Falkner, C. H., "Jointly Optimal Deterministic Inventory and Replacement Policies," Man. Sci.: Theory *16*, 622–635 (1970).

16 Fanarzhi, G. N. and D. V. Rozhadestvenskii, "Reliability of a Doubled System with Restoration and Preventive Maintenance Service," Izvestiya Akademii Nauk SSR, Tekhnicheskaya Kibernetika *3*, 61–66 (1970).

17 Faulkner, J. A., "The Use of Closed Queues in the Deployment of Coal Face Machinery," Oper. Res. Quart. *19*, 15–23 (1968).

18 Federowicz, A. J. and M. Mazumdar, "Use of Geometric Programming to Maximize Reliability Achieved by Redundancy," Oper. Res. *16*, 948–954 (1968).

19 Feeney, G. J. and C. C. Sherbrooke, "The $(s-1, s)$ Inventory Policy under Compound Poisson Demand," Man. Sci. *12*, 391–411 (1966).

20 Feldman, R. M., "Optimal Replacement for Shocks and Wear Processes," unpublished paper, Dept. of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Ill. (1974).

21 Feldman, R. M., "Optimal Replacement for Systems with Semi-Markovian Deterioration," Ph. D. Dissertation, Dept. of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Ill. (1975).

22 Fox, B., "Markov Renewal Programming by Linear Fractional Programming," SIAM J. Appl. Math. *14*, 1418–1432 (1966a).

[73] Fox, B., "Age Replacement with Discounting," Oper. Res. *14*, 533–537 (1966b).

[74] Fox, B., "Adaptive Age Replacement," J. Math. Anal. & Appl. *18*, 365–376 (1967).

[75] Fox, B., "Semi-Markov Processes: A Primer," The RAND Corporation, RM–5803 (1968).

[76] Gertsbakh, I. B., "Dynamic Reservation Optimal Control of Switch-In of Elements," Automatik i Vychislitel'naya Tekhnika *1*, 28–34 (1970).

[77] Glasser, G. J., "The Age Replacement Problem," Technometrics *9* (1967).

[78] Goss, T. I., "Truncated Sequential Test for the Variance of a Normal Distribution with App cations to Maintainability," AES–746, Institute of Admin. and Mgmt., Union College, Sch nectady, N.Y. (1974a).

[79] Goss, T. I., "Nonparametric Truncated Sequential Test for the Median with Application Maintainability," AES–745, Institute of Admin. and Mgmt., Union College, Schenectady, N.Y (1974b).

[80] Govil, A. K., "Stochastic Behavior of a Complex System with Bulk Failures and Priority Repairs, Rev. Frse. d'Informatique et de Recherche Operationnelle, *5*, No. V–1, 51–58 (1971).

[81] Greenberg, H., "An Application of a Lagrangian Penalty Function to Obtain Optimal Redu dancy," Technometrics *12*, 545–552 (1970).

[82] Greenberg, H. and W. Pierskalla, "A Review of Quasi-Convex Functions," Oper. Res. *19*, 1553 1570 (1971).

[83] Greenberg, H. and T. Robbins, "Finding Everett's Lagrange Multipliers by Generalized Linea Programming," Technical Report CP–70008, Parts I and II, Computer Science/Operation Research Center, Southern Methodist University, Dallas, Tex. (1972).

[84] Grinyer, P. H. and D. G. Toole, "A Note on the Theory of Replacement," INFOR *10*, 107–12 (1972).

[85] Hastings, N. A. J., "The Repair Limit Replacement Method," Oper. Res. Quart. *20*, 337–35 (1969).

[86] Henin, C., "Optimal Replacement Policies for a Single Loaded Sliding Standby," Man. Sci Theory *18*, 706–715 (1972).

[87] Henin, C., "Optimal Allocation of Unreliable Components for Maximizing Expected Profit ov Time," Nav. Res. Log. Quart. *20*, 395–403 (1973).

[88] Hinomoto, H., "Sequential Control of Homogeneous Activities—Linear Programming of Sem Markovian Decision," Oper. Res. *19*, 1664–1674 (1971).

[89] Hirsch, W., M. Meisner, and C. Boll, "Cannibalization in Multicomponent Systems and th Theory of Reliability," Nav. Res. Log. Quart. *15*, 331–359 (1968).

[90] Hochberg, M., "Generalized Multicomponent Systems under Cannibalization," Nav. Res. Lo Quart. *20*, 585–605 (1973).

[91] Hodgson, V. and T. L. Hebble, "Nonpreemptive Priorities in Machine Interference," Oper. Re *15*, 245–253 (1967).

[92] Hopkins, D., "Infinite-Horizon Optimality in an Equipment Replacement and Capacity E pansion Model," Man. Sci.: Theory *18*, 145–156 (1971).

[93] Howard, R., *Dynamic Programming and Markov Processes* (The MIT Press, Cambridge, Mass 1960).

[94] Howard, R., "System Analysis of Semi-Markov Processes," IEEE Transactions on Milita Electronics (Apr. 1964), pp. 114–124.

5] Howard, R., *Dynamic Probabilistic Systems* (John Wiley and Sons, New York, 1971).
6] Hsu, J. I. S., "An Empirical Study of Computer Maintenance Policies," Man. Sci. *15*, B180–B195 (1968).
7] Intriligator, M. D., *Mathematical Optimization and Economic Theory* (Prentice-Hall, Englewood Cliffs, N.J., 1971).
8] Jacobsen, S. E. and S. Arunkumar, "Investment in Series and Parallel Systems to Maximize Expected Life," Man. Sci. *19*, 1023–1028 (1973).
9] Jacobson, L. J., "Optimal Allocation of Resources in the Purchase of Spare Parts and/or Additional Service Channels," University of California at Berkeley, ORC 69–12 (1969).
0] Jain, A. and K. P. K. Nair, "Comparison of Replacement Strategies for Items that Fail," IEEE Transactions on Reliability *R—23*, 247–251 (1974).
1] Jain, J. P., "Overtime Concept in Provisioning of Spare Equipments," Opsearch *10*, 43–49 (1973).
2] Johnson, E., "Computation and Structure of Optimal Reset Policies," J. Am. Stat. Assoc. *62*, 1462–1487 (1967).
3] Jorgenson, D., J. McCall, and R. Radner, *Optimal Replacement Policies* (Rand McNally, 1967).
4] Kabak, I., "System Availability and Some Design Implications," Oper. Res. *17*, 827–837 (1969).
5] Kalman, P. J., "A Stochastic Constrained Optimal Replacement Model: The Case of Ship Replacement," Oper. Res. *20*, 327–334 (1972).
6] Kalymon, B. A., "Machine Replacement with Stochastic Costs," Man. Sci.: Theory *18*, 288–298 (1972).
7] Kamien, M. and N. Schwartz, "Optimal Maintenance and Sale Age for a Machine," Man. Sci. *17*, B495–B504 (1971).
] Kander, Z. and A. Raviv, "Maintenance Policies when Failure Distribution of Equipment is only Partially Known," Nav. Res. Log. Quart. *21*, 419–429 (1974).
] Kao, E., "Optimal Replacement Rules," Oper. Res. *21*, 1231–1249 (1973).
] Kaplan, S., "A Note on a Constrained Replacement Model for Ships Subject to Degradation by Utility," Nav. Res. Log. Quart. *21*, 563–568 (1974).
] Keller, J. B., "Optimum Checking Schedules for Systems Subject to Random Failure," Man. Sci. *21*, 256–260 (1974).
] Kent, A., "The Effect of Discounted Cash Flow on Replacement Analysis," Oper. Res. Quart. *21*, 113–118 (1960).
Klein, M., "Inspection-Maintenance-Replacement Schedules under Markovian Deterioration," Man. Sci. *9*, 25–32 (1962).
Klein, M. and R. Kirch, "Surveillance Schedules for Medical Examinations," Man. Sci. *20*, 1403–1409 (1974).
Kirsch, C. M., "A Method for Determining Bridge Painting Cycles," Materials Protection *4*, 40–49 (1966).
Koehler, G., A. Whinston, and G. Wright, "Matrix Iterative Techniques in Large Scale Linear Programming," Proceedings of IEEE Systems, Management and Cybernetics, Boston, Mass. (Nov. 1973a).

[117] Koehler, G., A. Whinston, and G. Wright, "The Solution of Leontieff Substitution Systems Us Matrix Iterative Techniques," Working Paper, Krannert School, Purdue University, W Lafayette, Ind. (1973b).

[118] Koehler, G., A. Whinston, and G. Wright, "An Iterative Procedure for Non-Discounted Discr Time Markov Decisions," Discussion Paper No. 70, Center for Math. Studies in Econ. Mgmt. Science, Northwestern University, Evanston, Ill. (1974).

[119] Kolesar, P., "Minimum Cost Replacement Under Markovian Deterioration," Man. Sci. 694–706 (1966).

[120] Kolesar, P., "Randomized Replacement Rules which Maximize the Expected Cycle Lengt Equipment Subject to Markovian Deterioration," Man. Sci. *13*, 867–876 (1967).

[121] Kolomenskii, L. V., "The Readiness Coefficient of Complex Systems," Automatika i Vyc litel'naya Tekhnika *2*, 62–65 (1970).

[122] Kopelevich, B. M., "Preventive Maintenance of Duplicate Systems," Automatika i Vychislitel'n Tekhnika *5*, 49–55 (1969).

[123] Kopelevich, B. M., "Preventive Maintenance on Complex Systems," Automatika i Vyc litel'naya Tekhnika *1*, 37–42 (1971).

[124] Korneichuk, V. I., "On the Effectiveness of Conditioning and Reservation of Elements," A matika i Telemekhanika *6*, 156–160 (1970).

[125] Kulshrestha, D. K., "Operational Behaviour of a Multicomponent System Having Stand-by dundancy with Opportunistic Repair," Unternehmensforschung *12*, 159–172 (1968a).

[126] Kulshrestha, D. K., "Reliability of a Parallel Redundant Complex System," Oper. Res. 28–35 (1968b).

[127] Kulshrestha, D. K., "Reliability of a Repairable Multicomponent System with Redundanc Parallel," IEEE Transactions on Reliability *R–19*, 50–53 (1970).

[128] Kumagai, M., "Reliability Analysis for Systems with Repair," J. Oper. Res. Soc., Japan 53–71 (1971).

[129] Lambe, T. A., "The Decision to Repair or Scrap a Machine," Oper. Res. Quart. *25*, 99-1 (1974).

[130] Lambert, B. K., A. G. Walvekar, and J. P. Hirmas, "Optimal Redundancy and Availab Allocation in Multistage Systems," IEEE Transactions on Reliability, *R–20*, 182–185 (197

[131] Lasdon, L. S., "A Survey of Large Scale Mathematical Programming," Tech. Memo. No. Dept. of Operations Research, Case-Western Reserve University, Cleveland, Ohio (1974).

[132] Lebedintseva, E. P., "Distribution of a Function of the Cost of Losses in Problems of Cor and Preventive Maintenance," Kibernetika *5*, 118–121 (1969).

[133] Liebling, T., G. Amiot, and X. Moreau, "Maintenance of a System with Elements that have Ra dom Life Times," New Techniques *13*, 55–60 (1971).

[134] Luss, H. and Z. Kander, "A Preparedness Model Dealing with *N* Systems Operating Simult ously," Oper. Res. *22*, 117–128 (1974).

[135] Luss, H. and Z. Kander, "Inspection Policies when Duration of Checkings is Non-Negligib Oper. Res. Quart. *25*, 299–309 (1974).

[136] Maaløe, E., "Approximate Formula for Estimation of Waiting-Time in Multi-Channel Queuing System," Man. Sci. *19*, 703–710 (1973).

7] Mann, S. H., "On the Optimal Size for Exploited Natural Animal Population," Oper. Res. *21*, 672–676 (1973).

8] Marshall, A. W. and F. Proschan, "Mean Life of Series and Parallel Systems," J. Appl. Prob. 7, 165–174 (1970).

9] Marshall, A. W. and F. Proschan, "Classes of Distributions Applicable in Replacement, with Renewal Theory Implications," Proc. Sixth Berkeley Symp. Math. Statist. Prob. I (1972), pp. 395–415.

0] Martz, H. F., Jr., "On Single Cycle Availability," IEEE Transactions on Reliability *R–20*, 21–23 (1971).

1] Masse, P., *Optimal Investment Decisions* (Prentice-Hall, Englewood Cliffs, N.J., 1962).

2] McCall, J. J., "Maintenance Policies for Stochastically Failing Equipment: A Survey," Man. Sci. *11*, 493–524 (1965).

3] McNichols, R. and G. Messer, "A Cost-Based Availability Allocation Algorithm," IEEE Transactions on Reliability *R–20*, 178–182 (1971).

4] Meisel, W. S., "On-Line Optimization of Maintenance and Verification Schedules," IEEE Transactions on Reliability *R–18*, 200–201 (1969).

5] Meyer, R. A., Jr., "Equipment Replacement under Uncertainty," Man. Sci.: Theory *17*, 750–758 (1971).

6] Mikami, M., "On Application of Erlang Distribution to Repairing a Group of Machines," Kyudai Kogaku Shuho (Technology Reports, Kyushu University) *44*, 1–8 (1971a).

7] Mikami, M., "On General Erlang Distribution Applied to Repairing of a Group of Machines—The Case of 2 Channels," Kyushu Daigaku Kogaku Shuho *44*, 131–138 (1971b).

8] Miller, B. L., "A Multi-Item Inventory Model with a Joint Backorder Criterion," Oper. Res. *19*, 1467–1476 (1971).

9] Miller, B. L., "Dispatching from Depot Repair in a Recoverable Item Inventory System: On the Optimality of a Heruistic Rule," Working paper, Western Management Science Institute, University of California, Los Angeles, Calif. (1973).

0] Mine, H. and H. Kawai, "Preventive Maintenance of a 1-Unit System with a Wearout State," IEEE Transactions on Reliability *R–23*, 24–29 (1974a).

1] Mine, H. and H. Kawai, "An Optimal Maintenance Policy for a 2-Unit Parallel System with Degraded State, IEEE Transactions on Reliability *R–23*, 81–86 (1974b).

2] Misra, K. B., "Reliability Optimization of a Series Parallel System," IEEE Transactions on Reliability *R–21*, 230–238 (1972).

3] Moore, J. R., Jr., "Forecasting and Scheduling for Past-Model Replacement Parts," Man. Sci.: Appl. *18*, B200–B213 (1971).

4] Moore, S. C., W. M. Faucett, R. D. Gilbert, and R. W. McMichael, "Computerized Selection of Aircraft Spares Inventories," Fort Worth Div. of General Dynamics, Fort Worth, Tex. (1970).

5] Morimura, H., "On Some Preventive Maintenance Policies for IFR," J. Oper. Res. Soc., Japan *12*, 94–124 (1970).

6] Morimura, H. and H. Makabe, "On Some Preventive Maintenance Policies," J. Oper. Res. Soc., Japan *6*, 17–47 (1963a).

[157] Morimura, H. and H. Makabe, "A New Policy for Preventive Maintenance," J. Oper. Res. Soc Japan *5*, 110–124 (1963b).

[158] Morimura, H. and H. Makabe, "Some Considerations on Preventive Maintenance Policies wit Numerical Analysis," J. Oper. Res. Soc., Japan *7*, 154–171 (1964).

[159] Moses, M., "Dispatching and Allocating Servers to Stochastically Failing Networks," Man. Sc *18*, B289–B300 (1972).

[160] Munford, A. G. and A. K. Shahani, "A Nearly Optimal Inspection Policy," Oper. Res. Quart. 2 373–379 (1972).

[161] Munford, A. G. and A. K. Shahani, "An Inspection Policy for the Weibull Case," Oper. Re Quart. *24*, 453–458 (1973).

[162] Nair, K. P. K. and M. D. Naik, "Multistage Replacement Strategies," Oper. Res. *13*, 279–29 (1965a).

[163] Nair, K. P. K. and M. D. Naik, "Multistage Replacement Strategies with Finite Duration Transfer," Oper. Res. *13*, 828–835 (1965b).

[164] Nair, K. P. K. and V. P. Marathe, "On Multistage Replacement Strategies," Oper. Res. *1* 537–539 (1966a).

[165] Nair, K. P. K. and V. P. Marathe, "Multistage Planned Replacement Strategies," Oper. Re *14*, 874–887 (1966b).

[166] Nakagawa, T., "The E [Number of Visits to State k Before a Total System Failure of a Comple System with Repair Maintenance]," Oper. Res. *22*, 108–111 (1974).

[167] Nakagawa, T. and S. Osaki, "Some Comments on a 2-Unit Standby Redundant System," Kei Kagaku *14*, 29–34 (1971).

[168] Nakagawa, T. and S. Osaki, "The Optimal Repair Limit Replacement Policies," Oper. Re Quart. *25*, 311–317 (1974).

[169] Nakamichi, H., J. Fukata, S. Takamatsu, and M. Kodoma, "Reliability Considerations on Repairable Multicomponent System with Redundancy in Parallel," J. Oper. Res. Soc., Japa *17* (1974).

[170] Naslund, B., "Simultaneous Determination of Optimal Repair Policy and Service Life," Swedis J. Economics, Vol. 68, No. 2, 63–73 (1966).

[171] Neuvians, G. and H. J. Zimmerman, "The Determination of Optimal Replacement and Maint nance Policies by Dynamic Programming," Ablauf und Planungsforschung *11*, 94–104 (1970

[172] Nicholson, T. A. J. and R. D. Pullen, "Dynamic Programming Applied to Ship Fleet Manag ment," Oper. Res. Quart. *22*, 211–220 (1971).

[173] Nylehn, B. and H. M. Blegen, "Organizing the Maintenance Function—An Analytical Approach International J. Production Res. *7*, 23–32 (1968).

[174] Onaga, K., "Maintenance and Operating Characteristics of Communication Networks," Ope Res. *17*, 311–336 (1969).

[175] Osaki, S., "Reliability Analysis of a 2-Unit Standby Redundant System with Standby Failure Opsearch *7*, 13–22 (1970a).

[176] Osaki, S., "System Reliability Analysis by Markov Renewal Processes," J. Oper. Res. Soc Japan *12*, 127–188 (1970b).

[177] Osaki, S., "Reliability Analysis of a Standby Redundant System with Preventive Maintenance Keiei Kagaku *14*, 233–245 (1971).

8] Osaki, S., "An Intermittently Used System with Preventive Maintenance," Keiei Kagaku *15*, 102–111 (1972).

9] Osaki, S. and Asakura, "A Two-Unit Standby Redundant System with Repair and Preventive Maintenance," J. Appl. Prob. *7*, 641–648 (1970).

0] Osaki, S. and T. Nakagawa, "Optimal Preventive Maintenance Policies for a 2-Unit Redundant System," IEEE Transactions on Reliability *R–23*, 86–91 (1974).

1] Polyak, D. G., "Calculation of the Reliability of Restorable Systems Taking Account of Work Shifts of Repair Personnel," Automatika i Telemekhanika *12*, 103–109 (1968).

2] Port, S. C., "Optimal Procedures for the Installation of a Unit Subject to Stochastic Failures," J. Math. Anal. & Appl. *9*, 491–497 (1964).

3] Port, S. C. and J. Folkman, "Optimal Procedures for Stochastically Failing Equipment," J. Appl. Prob. *3*, 521–537 (1966).

4] Porteus, E. and Z. Lansdowne, "Optimal Design of a Multi-Item Multi-Location Multi-Repair Type Repair and Supply System," Nav. Res. Log. Quart. *21*, 213–238 (1974).

5] Proschan, F., "Recent Research on Classes of Life Distributions Useful in Maintenance Modeling," Florida State University Statistics Report M291. AFOSR Tech. Report No. 26 (1974).

6] Proschan, F. and T. A. Bray, "Optimal Redundancy Under Multiple Constraints," Oper. Res. *13*, 800–814 (1965).

7] Quayle, N. J. T., "Damaged Vehicles—Repair or Replace," Oper. Res. Quart. *23*, 83–87 (1972).

8] Rangnekar, S. S., "Effect of Inflation on Equipment Replacement Policy," Opsearch *4*, 149–163 (1967).

9] Reinitz, R. C. and L. Karasyk, "A Stochastic Model for Planning Maintenance of Multi-part Systems," from *Proc. of the Fifth International Conf. on O. R.*, J. Lawrence, ed. (1969) 703–713.

0] Roeloffs, R., "Minimax Surveillance Schedules with Partial Information," Nav. Res. Log. Quart. *10*, 307–322 (1963).

1] Roeloffs, R., "Minimax Surveillance Schedules for Replaceable Units," Nav. Res. Log. Quart. *14*, 461–471 (1967).

2] Rolfe, A. J., "Markov Chain Analysis of a Situation where Cannibalization is the Only Repair Activity," Nav. Res. Log. Quart. *17*, 151–158 (1970).

3] Roll, Y. and P. Naor, "Preventive Maintenance of Equipment Subject to Continuous Deterioration and Stochastic Failure," Oper. Res. Quart. *19*, 61–73 (1968).

4] Rose, M., "An Investment Model for Repairable Assets: The $F-Y$ Case," Research Contribution 31, Institute of Naval Studies, Center for Naval Analysis, Arlington, Va. (1969).

5] Rose, M., "Computing the Expected End-Product Service Time Using Stochastic Item Delays," Oper. Res. *19*, 524–540 (1971).

6] Rose, M., "Determination of the Optimal Investment in End Products and Repair Resources," Nav. Res. Log. Quart. *20*, 147–159 (1973).

7] Rose, M. and L. Brown, "An Incremental Production Function for the End-Item Repair Process," AIIE Trans. *II*, 166–171 (1970).

8] Ross, S., "Non-Discounted Denumerable Markovian Decision Models," Annals of Math. Statist. *39*, 412–423 (1968a).

9] Ross, S., "Denumerable Markovian Decision Models," Ph. D. Dissertation, Department of Statistics, Stanford University, Stanford, Calif. (1968b).

[200] Ross, S., "A Markovian Replacement Model with a Generalization to Include Stocking," Ma Sci.: Theory *15*, 702–715 (1969a).

[201] Ross, S., "Average Cost Semi-Markov Decision Processes," Oper. Res. Center, University California at Berkeley, ORC 69–27, Berkeley, Calif. (1969b).

[202] Ross, S., *Applied Probability Models with Optimization Applications* (Holden Day, San Francisc Calif. (1970).

[203] Ross, S., "On Time to First Failure in Multicomponent Exponential Reliability Systems," Ope Res. Center, University of California at Berkeley, ORC 74–8, Berkeley, Calif. (1974).

[204] Ruben, R. V., "The Economically Optimal Period of Preventive Maintenance of a System wi Possible Dislocations," Automatika i Vychislitel'naya Tekhnika *2*, 30–34 (1971).

[205] Sackrowitz, H. and E. Samuel-Cahn, "Inspection Procedures for Markov Chains," Man. S *21*, 261–270 (1974).

[206] Satia, J., "Markovian Decision Processes with Uncertain Transition Matrices and/or Probabilist Observation of States," unpublished Ph. D. Dissertation, Stanford University, Stanford, Cal (1968).

[207] Satia, J. and R. Lave, "Markovian Decision Processes with Uncertain Transition Probabilities Oper. Res. *21*, 728–740 (1973).

[208] Savage, R., "Surveillance Problem," Nav. Res. Log. Quart. *9*, 187–209 (1962).

[209] Savage, R. and G. R. Antelman, "Surveillance Problems: Weiner Processes," Nav. Res. Lo Quart. *12*, 35–55 (1965).

[210] Scheaffer, R. L., "Optimum Age Replacement Policies with an Increasing Cost Factor," Techn metrics *13*, 139–144 (1971).

[211] Schrady, D. A., "A Deterministic Inventory Model for Repairable Items," Nav. Res. Log. Qua *14*, 391–398 (1967).

[212] Schwartz, A., J. Sheler, and C. Cooper, "Dynamic Programming Approach to the Optimizati of Naval Aircraft Rework and Replacement Policies," Nav. Res. Log. Quart. *18*, 395–414 (197

[213] Schweitzer, P. J., "Optimal Replacement Policies for Hyperexponentially and Uniformly D tributed Lifetimes," Oper. Res. *15*, 360–362 (1967).

[214] Scott, M., "On the Number of Tasks by a Machine Subject to Breakdown," Cah. Cent. d'E Rech. Oper. *9*, 44–56 (1967).

[215] Scott, M., "A Problem in Machine Breakdown," Unternehmensforschung *12*, 23–33 (196

[216] Scott, M., "Distribution of the Number of Tasks by a Repairable Machine," Oper. Res. *2* 851–859 (1972).

[217] Serfozo, R., "A Replacement Problem Using a Wald Identity for Discounted Variables," Ma Sci. *20*, 1314–1315 (1974).

[218] Serfozo, R. and R. Deb, "Optimal Control of Batch Service Queues," Adv. Appl. Prob. *5*, 340–3 (1973).

[219] Sethi, S., "Simultaneous Optimization of Preventive Maintenance and Replacement Policies f Machines: A Modern Control Theory Approach," AIIE Transactions *5*, 156–163 (1973).

[220] Sethi, S. and T. E. Morton, "A Mixed Optimal Technique for Generalized Machine Replac ment Problem," Nav. Res. Log. Quart. *19*, 471–482 (1972).

[221] Sherbrooke, C. C., "METRIC: A Multi-Echelon Technique for Recoverable Item Control Oper. Res. *16*, 122–141 (1968).

Sherbrooke, C. C., "An Evaluation for the Number of Operationally Ready Aircraft in a Multilevel Supply System," Oper. Res. *19*, 618–635 (1971).

Shershin, A., "Mathematical Optimization Techniques for the Simultaneous Apportionment of Reliability and Maintainability," Oper. Res. *18*, 95–106 (1970).

Silver, E., "Inventory Allocation Among an Assembly and Its Repairable Subassemblies," Nav. Res. Log. Quart. *19*, 261–280 (1972).

Simon, R. M., "The Reliability of Multistate Systems Subject to Cannibalization," Report AM–69–1, School of Engineering and Applied Science, Washington University, St. Louis (1969).

Simon, R. M., "Optimal Cannibalization Policies for Multicomponent Systems, SIAM J. Appl. Math. *19*, 700–711 (1970).

Simon, R. M., "The Reliability of Multicomponent Systems Subject to Cannibalization," Nav. Res. Log. Quart. *19*, 1–14 (1972).

Simon, R. M. and D. A. D'Esopo, "Comments on a Paper by S. G. Allen and D. A. D'Esopo: 'An Ordering Policy for Repairable Stock Items'," Oper. Res. *19*, 986–989 (1971).

Sivazlian, B. D., "On a Discounted Replacement Problem with Arbitrary Repair Time Distribution," Man. Sci. *19*, 1301–1309 (1973).

Smallwood, R. and E. Sondik, "The Optimal Control of Partially Observable Markov Processes over a Finite Horizon," Oper. Res. *21*, 1071–1088 (1973).

Sobel, M. J., "Production Smoothing with Stochastic Demand and Related Inventory Problems," Tech. Report No. 17, Dept. of Operations Research and the Dept. of Statistics, Stanford University, Stanford, Calif. (1967).

Soland, R. M., "A Renewal Theoretic Approach to the Estimation of Future Demand for Replacement Parts," Oper. Res. *16*, 36–51 (1968).

Solov'ev, A. D., "Reservation with Rapid Recovery," Izvestiya Akademii Nauk SSSR. Tekhnicheskaya Kibernetika, No. 1, 56–71 (1970).

Sondik, E., "The Optimal Control of Partially Observable Markov Processes," Ph. D. Dissertation, Dept. of Eng. Econ. Systems, Stanford University (1971).

Srinivasan, S. K., "The Effect of Standby Redundancy in System's Failure with Repair Maintenance," Oper. Res. *14*, 1024–1036 (1966).

Srinivasan, S. K., "First Emptiness in the Spare Parts Problem for Repairable Components," Oper. Res. *16*, 407–415 (1968).

Srinivasan, S. K. and M. Gopalan, "Probabilistic Analysis of a 2-Unit System with a Warm Standby and a Single Repair Facility," Oper. Res. *21*, 748–754 (1973).

Tahara, A. and T. Nishida, "Optimal Replacement Policies for a Repairable System with Markovian Transition of State," J. Oper. Res. Soc., Japan *16*, 78–103 (1973).

Tapiero, C. S., "Optimal Simultaneous Replacement and Maintenance of a Machine with Process Discontinuities," Rev. Frse. d'Informatique et Recherche Operationnelle, No. 2, 79–86 (1971).

Tapiero, C. S., "Optimal Maintenance and Replacement of a Sequence of *n*-Machines and Technical Obsolescence," Opsearch *10*, 1–13 (1973).

Taylor, H. M., "Optimal Replacement Under Additive Damage and Other Failure Models," unpublished paper, Oper. Res. Dept., Cornell University, Ithaca, N.Y. (1973).

Taylor, J. and R. R. P. Jackson, "Application of Birth and Death Processes to Provisions to Spare Machines," Oper. Res. Quart. *5*, 95–108 (1954).

[243] Thompson, G. L., "Optimal Maintenance Policy and Sale Date of a Machine," Man. Sci.: Theo *14*, 543–550 (1968).

[244] Tillman, F. A., "Optimization by Integer Programming of Constrained Reliability Proble with Several Modes of Failure," IEEE Transactions on Reliability *R–18*, 47–53 (1969).

[245] Tillman, F. A. and J. M. Liittschwager, "Integer Programming Formulation to Constraine Reliability Problems," Man. Sci. *13*, 887–899 (1967).

[246] Turban, E., "The Use of Mathematical Models in Plant Maintenance Decision Making," Ma Sci. *13*, 342–358 (1967).

[247] Veinott, A. F., Jr., "Optimal Policy for a Multi-Product, Dynamic, Nonstationary Invento Problem," Man. Sci. *12*, 206–222 (1965).

[248] Veinott, A. F., Jr., "Extreme Points of Leontieff Substitution Systems," Linear Algebra and I Applications *1*, 181–194 (1968).

[249] Vergin, R. C., "Scheduling Maintenance and Determining Crew Size for Stochastically Failin Equipment," Man. Sci. *12*, B52–B65 (1966).

[250] Vergin, R. C., "Optimal Renewal Policies for Complex Systems," Nav. Res. Log. Quart. *1*, 523–534 (1968).

[251] von Ellenrieder, A. and A. Levine, "The Prob. of an Excessive Nonfunctioning Interval," Ope Res. *14*, 835–840 (1966).

[252] Washburn, L. A., "Determination of Optimum Burn-In Time: A Composite Criterion," "IEE Transactions on Reliability *R–19*, 134–140 (1970).

[253] Weiss, G., "On Some Economic Factors Influencing a Reliability Program," NAVORD 425 Naval Ordinance Laboratory, White Oak, Md. (1956).

[254] Wolff, M. R., "Optimale Instandhaltungspolitiken in Einfachen Systemen," Lecture Notes i Oper. Res. *18*, Springer, Berlin (1970).

[255] Wolff, M. R. and R. Subramanian, "Optimal Readjustment Intervals," Oper. Res. *22*, 191–19 (1974).

[256] Woodman, R. C., "Replacement Policies for Components that Deteriorate," Oper. Res. Quar *18*, 267–281 (1967).

[257] Wright, G. P. and J. Prawda, "On a Replacement Problem," Cah. Cent. d'Et. Rech. Oper. *1* 43–52 (1972).

[258] Zachs, S. and W. Fenske, "Sequential Determination of Inspection Epochs for Reliability Sy tems with General Lifetime Distributions," Nav. Res. Log. Quart. *20*, 377–386 (1973).

[259] Zak, Yu. I., "A Problem of Determination of the Optimal Sequencing of Readjustments of Equi ment," Kibernetika, *6*, 86–92 (1968).

MARKOVIAN DETERIORATION WITH UNCERTAIN INFORMATION—A MORE GENERAL MODEL

Donald Rosenfield
State University of New York
at Stony Brook
Stony Brook, N.Y.

ABSTRACT

A model of a deteriorating system with imperfect information is considered. The structures appropriate for such a model include failing machinery and depleted inventory systems. In an effort to add a new dimension to such models, it is assumed that the operator must pay an inspection cost to determine the precise state of the system. At the start of every time period, the operator is faced with three choices: repair, no action, or inspection. Under fairly general assumptions, the optimal policy for repair is found to be straightforward and intuitive. This result has two important areas of application.

INTRODUCTION

This paper is concerned with finding the form of optimal policies for deteriorating Markov processes imperfect information, i.e., where the state is partially observable. If the state is partially observable, then the operator has some information about the state, but he cannot precisely determine state. Part of his task is determination of when an inspection cost should be paid to determine the e.

Examples of such processes include deteriorating machinery or an inventory system being depleted. achine, for example, could fail or deteriorate to the point where large losses of production occur. nventory systems, processes deterioriate in the sense that stock becomes depleted, resulting in possibility of unfilled orders. In both of these cases, the manager must decide on an inspection and ir policy to minimize his losses.

In the model presented, the process undergoes deterioration according to a discrete-time Markov ess, and operating and repair costs are presumed to increase with state number.

Models such as those in [2] and [7] analyzed such processes, and assume that the operator has ect knowledge of the state of the system. Under such an assumption, the operator is concerned the optimal times to repair the machine and the optimal times to take no action. In other words, r the perfect information assumption, the operator must decide between two actions. Under the mption that the state is partially observable, the operator has a choice of three actions: repair, ction, or inspection. Upon inspection, the operator pays an inspection cost and determines the ise state of the system. It is felt that the assumption of imperfect information adds a different nsion to the general problem.

Following models in [10], [17] and [18], the author [15] introduced a general model that incorporated ssumption of imperfect information. The author also obtained definitive policy results for the model.

The present paper is a followup to that paper in that it presents weaker results under weaker assump tions on the transition matrix.

The result of this paper states that the optimal region for repair is of a certain intuitive type (se Figure 1). The formal definition of such a policy is presented in section 1.

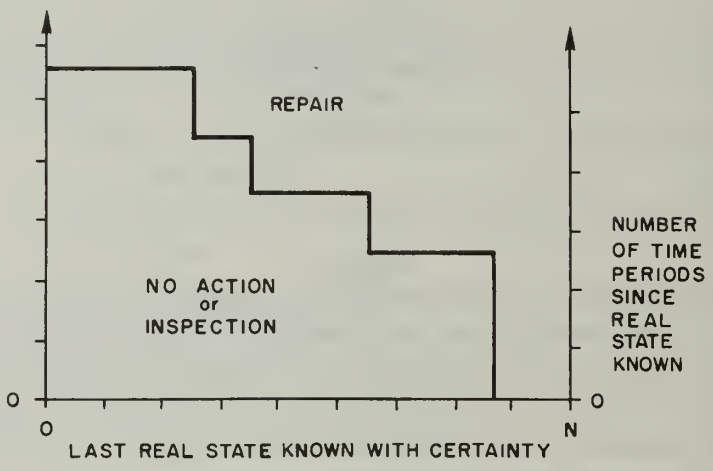


FIGURE 1. A monotonic and optimal policy.

1. MODEL AND RESULTS

We have an underlying discrete-time Markov process, representing, for example, a deterioratin machine or an inventory system undergoing depletion. The process is characterized by the state (referred to as real states) 0,1, . . ., N, and transition matrix P. Under the structure that we have intro duced, when the process is in "observed" state (i, k), then the operator knows that k time units ag the process was in real state i. Furthermore, the operator has not obtained any additional informatio in the past k time units. Under the repair option, the operator assumes an expected repair cost

$$\sum_{j=0}^{N} p_{ij}^{k} C_j,$$

where C_j, $j=0, \ldots, N$ is the repair cost for real state j, and the process reverts to state (0, 0). (W are using p_{ij}^{k} as the element ij in the matrix P^k. This value is equal to the a posteriori probability tha the real state is j given an observed state (i, k).) Under the other two options, the operator assume an expected operating cost

$$\sum_{j=o}^{N} p_{ij}^{k} L_j,$$

where L_j, $j=0, 1, \ldots, N$ is the one-period operating cost for state j. Under the inspection option, th operator assumes also an inspection cost M, and the process reverts to state $(j, 0)$ at the end of the cu rent time period with probability p_{ij}^{k+1}. Note that inspection is perfect in that it gives the operator th precise state. Under the final option (no action), the process undergoes transition to state $(i, k+1)$. I is seen that the space (i, k), $i=0, 1, \ldots, N$, $k \geq 0$ is sufficient to describe the operator's knowledge o the process.

The model of Derman [2] was an important influence on this paper and its predecessor. Derman esented a multistate model with perfect information and proved the existence of an intuitive optimal licy. Kolesar [11] extended the result. Girshick and Rubin [7] and Taylor [20] looked at two-state odels where an observable quality characteristic is recorded each time period. Girshick and Rubin lved this problem, but incorrectly hypothesize optimal policy when observation of the quality char- teristic is a costly option. Taylor found a counterexample to their hypothesis. Such models with three tions (repair, inspection, or no action) have been examined by Smallwood and Sondik [18], who ovide a computational algorithm for solving the finite-horizon case, Klein [10], who formulates the oblem as a linear program, Emmons [5], who introduces a model similar to this author's, and Ross 7], who presents a generalized model and obtains definitive policy results for the case of two real ates. In the recent paper of this author [15], the general model of the present paper was first set th. The author also found definitive policy results for the model. Emmons [5] showed the optimal gion for repair as i varies for case $k=0$, but this result is less inclusive than the results that follow.

We now define a monotonic policy (Figure 1):

DEFINITION: Suppose a policy is defined by numbers $k^*(i)$ nonincreasing in i such that for state k) a repair is made if $k \geq k^*(i)$, and no action or inspection is called for otherwise. Then we call e policy a monotonic policy and the numbers $k^*(i)$ critical numbers.

The major result of this paper is that a monotonic policy is optimal under fairly general assump- ns. This optimality holds under both the discounted and average-cost criteria. It is felt that a mono- ic policy is an intuitive one.

In reference to this result, some comparison should be made to the results of the previous paper of author [15]. There, it is shown that a monotonic policy is optimal under stronger assumptions. In dition, the optimal region for inspection is also shown. A monotonic policy is concerned only with optimal region for repair. The optimality of the detailed inspection structure of the previous paper not been shown under the new assumptions, except for a subcase that we discuss later. The optimal- of a monotonic policy, however, is shown under much more general assumptions than in [15].

The assumptions that we make are:

C_j and L_j are nondecreasing in j

$C_j - L_j$ is nonincreasing in j

There is a number $\delta > 0$ and an integer r such that $\min_{0 \leq i \leq N} p_{ij}^r \geq \delta$ for any recurrent state j.

a) $\sum_{j=k}^{N} p_{ij}$ is nondecreasing in i, $k=0, 1, \ldots, N$ (increasing failure rate, or IFR) and

b) for any vector $F=\{F_j,\ j=0,\ 1,\ \ldots,\ N\colon F_j$ is nonincreasing in $j\}$, $(P \cdot F)_i \leq F_i V(\Pi \cdot F)$, $i=1, \ldots, N-1$, where Π is the vector of limiting state probabilities and V denotes the maximum operator.

Assumptions 1 and 2 are identical to the corresponding assumptions in [15] and formalize the ideas t the higher numbered states are costlier and that repair improves relative to no action as the state nber increases. Assumption 3 is analogous to Assumption 3 of [15] and insures one ergodic class. only real difference is the matrix assumption, Assumption 4, which is a weaker assumption than umption 4 of [15]. The first part of the Assumption, the IFR assumption, is an intuitive notion eloped by Derman [2] that implies that the mass tends to the higher numbered states. The second

part of Assumption 4 is a new concept developed by the author. The condition can be checked by solving $N-1$ single constraint linear programs (see page 91 of [14]). Note in particular that the condition is trivially satisfied for a process with two real states ($N=1$), and is also satisfied if P is upper triangular. The two-state process is a special case that we shall examine later.

Under these assumptions, the optimal policy is often a monotonic policy. The determining conditions can be expressed in terms of the costs. In Theorem 3.1.2 and Corollary 3.1.1 of the author's technical report [14], conditions for trivial (perpetual no action) optimal policies are given. Assume the optimal policy is not the trivial one. Then the major theoretical result of this paper is: If the limiting optimal action in state (N, k) as $k \to \infty$ is repair, an assumption dependent on the costs C_j, L_j, and M then the optimal policy is a monotonic one. We shall present sufficient conditions on the costs for insuring that the limiting optimal action in state (N, k) is indeed repair. The result is of particular importance for two special cases.

2. THE DISCOUNTED CASE

We first present the mathematical definition relating to the concept of limiting action in (i, k) as $k \to \infty$. We define F_{ik} as the difference between the discounted cost choosing repair initially and the discounted cost using no action initially and subsequently following an optimal policy in both cases. We denote the discount factor as α and note that $0 \leq \alpha < 1$. We define G_{ik}, similarly, as the cost of repair minus the cost of inspection. Both differences are functions of the present state (i, k), and their signs determine optimal action at (i, k).

Given the recursive expression for $C(i, k)$, the optimal discounted cost over an infinite horizon when starting in state (i, k), we can determine the values of F_{ik} and G_{ik}. These values are obtained by examining the terms of the expression for $C(i, k)$:

$$C(i, k) = \text{Min}\left\{\sum_{j=0}^{N} p_{ij}^{k} C_j + \alpha C(0, 0),\ \sum_{j=0}^{N} p_{ij}^{k} L_j + M + \alpha \sum_{j=0}^{N} p_{ij}^{k+1} C(j, 0),\ \sum_{j=0}^{N} p_{ij}^{k} L_j + \alpha C(i, k+1)\right\}.$$

The three components in the minimization correspond to the choices of repair, inspection, and no action and $0 \leq \alpha < 1$ is the discount factor. Recursive expressions for optimal discounted cost are due to Taylor [20] and depend on the existence of stationary nonrandom optimal policies, which is due to Blackwell [1] and Derman [3].

By examining the expression, we see that

$$F_{ik} = \sum_{j=0}^{N} p_{ij}^{k} (C_j - L_j) + \alpha (C(0, 0) - C(i, k+1))$$

and

$$G_{ik} = \sum_{j=0}^{N} p_{ij}^{k} (C_j - L_j) - M + \alpha C(0, 0) - \alpha \sum_{j=0}^{N} p_{ij}^{k+1} C(j, 0).$$

We analogously denote $C^n(i, k)$ as the optimal discounted cost over a finite horizon. An analogous expression exists for the $C^n(i, k)$ expressed recursively in terms of $C^{n-1}(i, k)$. From Taylor [20], we

e the geometric convergence of $C^n(i, k)$ to $C(i, k)$. From the ergodicity assumption, it follows ictively that $C^n(i, k)$ has a limit as $k \to \infty$ (denoted by C^n_*) that is identical for all i. From the geo- ric convergence of $C^n(i, k)$ to $C(i, k)$ (and hence uniform convergence), we see that $C(i, k)$ con- ges to $C_* = \lim_{n\to\infty} C^n_*$ as $k \to \infty$. Consequently,

$$F_* = \lim_{k\to\infty} F_{ik} = \sum_{j=0}^{N} \Pi_j (C_j - L_j) + \alpha C(0, 0) - \alpha C_*$$

$$G_* = \lim_{k\to\infty} G_{ik} = \sum_{j=0}^{N} \Pi_j (C_j - L_j) + \alpha C(0, 0) - M - \alpha \sum_{j=0}^{N} \Pi_j C(j, 0).$$

en we say that the limiting optimal action in (N, k) (or (i, k)) as $k \to \infty$ is repair, we mean that ≤ 0 and $G_* \leq 0$. The concept is important because we use the limiting action of repair to prove the or result.

Note that the validity of these conditions $F_* \leq 0$, $G_* \leq 0$ depend on the values of $\{C(j, 0)\}$, $C = \{C_j\}$, $L = \{L_j\}$, and M. We assume $F_* \leq 0$ and $G_* \leq 0$ to obtain our result. Sufficient condi- s for these on C, L, and M, will be presented at the end of the section.

We precede the major result by some lemmas and the following definition. If $B_k \to B_\infty$ is a con- ent sequence, then we say that B_k satisfies property A when

$$B_{k+1} \leq \max (B_k, \lim B_k).$$

LEMMA 1: Under the assumptions, if G_j is nonincreasing, then

$$H_{ik} = \sum_{j=0}^{N} p_{ij}^k G_j$$

fies property A in k and is nonincreasing in i.

PROOF: Since P IFR $\to P^k$ IFR, the proof of monotonicity in i is due to Derman [2] (Lemma, p.

$$H_{0k} \geq \sum_{j=0}^{N} \Pi_j H_{jk} \qquad \text{since } H_{jk} \text{ is nonincreasing in } j.$$

$$= \sum_{j=0}^{N} \Pi_j G_j \qquad \text{by definition of } H_{jk} \text{ and by changing the order of summation,}$$

$$H_{0,k+1} = \sum_{j=0}^{N} p_{0j} H_{jk} \leq H_{0k} \text{ since } H_{jk} \text{ is nonincreasing in } j.$$

larly,

$$H_{Nk} \leq \sum_{j=0}^{N} \Pi_j H_{jk} = \sum_{j=0}^{N} \Pi_j G_j,$$

and

$$H_{N,k+1} \leq \sum_{j=0}^{N} \Pi_j H_{j,k+1} = \sum_{j=0}^{N} \Pi_j G_j.$$

Finally, for $1 \leq i \leq N-1$,

$$H_{i,k+1} = \sum_{j=0}^{N} p_{ij} H_{jk} \leq H_{ik} V\left(\sum_{j=0}^{N} \Pi_j H_{jk}\right) = H_{ik} V\left(\sum_{j=0}^{N} \Pi_j G_j\right)$$

since H_{ik} is nonincreasing in i and P satisfies Assumption 4.

LEMMA 2:

(a) Let $F_k = \{\underset{j \epsilon B}{\text{Max}}\ B_{jk}\}$ where B is a finite set and B_{jk} satisfies property A in k. Then F_k satisfies property A in k.

(b) Let F_k^n be a sequence of functions that satisfy Property A in k and denote $F_\infty^n = \lim_{k \to \infty} F_k^n$. Assume that $F_k^\infty = \lim_{n \to \infty} F_k^n$ exists and that

$$\lim_{k \to \infty} F_k^\infty = \lim_{n \to \infty} F_\infty^n.$$

Then F_k^∞ satisfies property A in k.

PROOF:

(a) Notice that

$$\lim_{k \to \infty} F_k = \max_{j \epsilon B} \{\lim_{k \to \infty} B_{jk}\}$$

so that

$$F_{k+1} = \underset{j}{\text{Max}}\ B_{j,k+1} \leq \underset{j}{\text{Max}}\ \text{Max}\ \{B_{jk}, \lim_{k \to \infty} B_{jk}\} \text{ by property } A$$

$$\leq \text{Max}\ \{\underset{j}{\text{Max}}\ B_{jk}, \underset{j}{\text{Max}} \lim_{k \to \infty} B_{jk}\}$$

$$= \text{Max}\ \{F_k, \lim_{k \to \infty} F_k\};$$

(b) $F_{k+1}^\infty = \lim_{n \to \infty} F_{k+1}^n \leq \lim_{n \to \infty} \text{Max}\ \{F_k^n, F_\infty^n\}$ by property A

$$= \text{Max}\ \{\lim_{n \to \infty} F_k^n, \lim_{n \to \infty} F_\infty^n\}$$

$$= \text{Max}\ (F_k^\infty, \lim_{k \to \infty} F_k^\infty).$$

are now ready to prove:

THEOREM 1: Assume $F_* \leqslant 0$ and $G_* \leqslant 0$. Then a stationary, monotonic policy is optimal for the ounted problem. Furthermore, it is optimal to repair in state (N, k), all k.

PROOF: We first assert that

$$C^n(i, k) = \min_{m \in [1, \bar{m}(n)]} \left\{ \sum_{j=0}^{N} p_{ij}^k B_{mj}^n \right\},$$

e $\bar{m}(n)$ is some finite value and B_{mj}^n is some nondecreasing function in j.

We show (1) by induction. (1) is trivial for $n = 1$. For case n,

$$C^n(i, k) = \text{Min} \left\{ \sum_{j=0}^{N} p_{ij}^k C_j + \alpha C^{n-1}(0, 0), \alpha \sum_{j=0}^{N} p_{ij}^k L_j + M + \right.$$

$$\left. \alpha \sum_{j=0}^{N} p_{ij}^{k+1} C^{n-1}(j, 0), \sum_{j=0}^{N} p_{ij}^k L_j + \alpha C^{n-1}(i, k+1). \right\}$$

$$\sum_{j=0}^{N} p_{ij}^k L_j + \alpha C^{n-1}(i, k+1) = \underset{m \in [1, \bar{m}(n-1)]}{\text{Min}} \left(\sum_{j=0}^{N} p_{ij}^k \left(L_j + \alpha \sum_{l=0}^{N} p_{jl} B_{ml}^{n-1} \right) \right)$$

$$\sum_{j=0}^{N} p_{ij}^{k+1} C^{n-1}(j, 0) = \sum_{j=0}^{N} p_{ij}^k \left(\sum_{l=0}^{N} p_{jl} C^{n-1}(l, 0) \right).$$

tituting these into (2) gives (1).

Also, by induction on n and Lemma 1,

$$C^n(i, k) \text{ is nondecreasing in } i,$$

hus

$$C(i, k) \text{ is nondecreasing in } i.$$

$$G_{ik} = \sum_{j=0}^{N} p_{ij}^k \left[C_j - L_j + \alpha \left(C(0, 0) - \sum_{j=0}^{N} p_{jl} C(l, 0) \right) \right]$$

ı increasing in i and satisfies Property A in k by Lemma 1.

Also,

$$F_{ik} = \lim_{n\to\infty} F^n_{ik} = \lim_{n\to\infty} \{p^k_{ij}(C_j - L_j) + \alpha(C^{n-1}(0,0) - C^{n-1}(C, k+1))\}$$

$$= \lim_{n\to\infty} \operatorname*{Max}_{m\in[1,\ldots,\bar{m}(n-1)]} \left\{ \sum_{j=0}^{N} p^k_{ij}(C_j - L_j + \alpha C^{n-1}(0,0) \right.$$

$$\left. - \alpha \sum_{l=0}^{N} p_{jl} B^{n-1}_{ml} \right) \Bigg\} \text{ by (1).}$$

Now

$$\lim_{k\to\infty} F^n_{ik} = \sum_{j=0}^{N} \Pi_j(C_j - L_j) + \alpha(C^{n-1}(0,0) - C^{n-1}_*).$$

Since

$$\lim_{k\to\infty} F_{ik} = \lim_{n\to\infty} \lim_{k\to\infty} F^n_{ik},$$

then by Lemmas 1 and 2, $C(i, k)$ is nonincreasing in i and satisfies property A in k. Now pick any i. By property A, as k is varied, F_{ik} is nonincreasing until F_{ik} goes below F_*. Beyond that point

$$F_{ik} \leq F_* \leq 0.$$

Thus, there is a

$$\underline{k} = \text{Min}\,\{k : F_{ik} \leq 0\},$$

such that for $\underline{k} \geq k$, $F_{ik} \leq 0$, and thus repair is just as good or better than no action. An analagous valu can be found for G_{ik}. Thus we can find critical numbers $k^*(i)$, such that repair is optimal for $k \geq k^*(i)$. The monotonicity of $k^*(i)$ follows directly from the monotonicity of F_{ik} and G_{ik} in i. Finally,

$$C^1(N, 1) = \text{Min} \left\{ \sum_{j=0}^{N} p_{Nj} C_j, \sum_{j=0}^{N} p_{Nj} L_j \right\} \geq \text{Min} \left\{ \sum_{j=0}^{N} p_{Nj} \sum_{j=0}^{N} p^m_{jl} C_l, \right.$$

$$\left. \sum_{j=0}^{N} p_{Nj} \sum_{j=0}^{N} p^m_{jl} L_l \right\} \text{ by Lemma 1}$$

$$= C^1(N, m+1).$$

$$C^1(N, 1) \geq \lim_{k \to \infty} C^1(N, k) = C^1_*.$$

induction and carrying to the limit,

$$C(N, 1) \geq C_*.$$

$$F_{N0} = C_N - L_N + \alpha C(0, 0) - \alpha C(N, 1) \leq F_*$$

6) and Assumption 2.

Similarly by (5) and Assumption 2,

$$G_{N0} \leq G_*.$$

roperty A, $F_{Nk} \leq F_* \leq 0$, $G_{Nk} \leq G_* \leq 0$, and repair is optimal in state (N, k), any k. □

Sufficient conditions for $F_* \leq 0$ and $G_* \leq 0$ and consequently the optimality of a nontrivial optimal cy are:

$$J_\alpha = \sum_{j=0}^{N} \Pi_j (C_j - L_j) + \sum_{i=1}^{\infty} \alpha^i \left[\sum_{j=0}^{N} p_{0j}^{i-1} L_j - \sum_{j=0}^{N} \Pi_j L_j \right] \leq 0,$$

$$K_\alpha = -M + \sum_{j=0}^{N} \Pi_j (C_j - L_j) + \alpha \max_{i^* \leq N} \sum_{j=0}^{N} \Pi_j D^\alpha_{i^* j} \leq 0,$$

re

$$D^\alpha_{i^* j} = \begin{cases} \operatorname{Min} \left[0, (1-\alpha) \sum_{i=0}^{\infty} \alpha^i \sum_{l=0}^{N} p^i_{0l} L_l - C_j \right] & j \geq i^* \\ L_0 - L_j + \alpha \sum_{l=0}^{N} (p_{0l} - p_{jl}) C_l & j < i^*. \end{cases}$$

amining these conditions, if

$$\sum_{j=0}^{N} \Pi_j (C_j - L_j) \leq 0,$$

the conditions are automatically satisfied. If not, the other nonpositive terms are added until hopefully the sum becomes nonpositive. $\sum_{j=0}^{N} \Pi_j(C_j - L_j)$ represents one-period cost differences inherent in F_* and G_*, and other terms represent bounds on future cost differences. A proof that these conditions are sufficient is presented in [14], Theorem 5.3.3 and Lemma 5.3.4.

3. THE AVERAGE-COST CASE

The results for the discounted case can be extended to give results for the average-cost case.

THEOREM 2: Assuming that the limiting optimal action in state (N, k) is repair for the average cost case, and, in particular, that

$$\lim_{\alpha_0 \to 1} \sup_{\alpha \geq \alpha_0} F_* < 0 \tag{7}$$

and

$$\lim_{\alpha_0 \to 1} \sup_{\alpha \geq \alpha_0} G_* < 0, \tag{8}$$

where it is noted that F_* and G_* depend on the discount factor α, then a monotonic policy is optimal under the average-cost problem, and it is optimal to repair in state (N, k) any k. Furthermore, all critical numbers are finite.

The extension of the discounted case to the average-cost case is generally a direct result of Ross [16], Theorems 1.1 to 1.3. Our proof extends these results since we also show the existence of finite critical numbers when (7) and (8) hold.

PROOF: Note that

$$C(i, k) \leq \sum_{j=0}^{N} p_{ij}^k C_j + \alpha C(0,0) \leq C_N + C(0,0),$$

where $C_N = \max_j \{C_j\} = C_j$ evaluated at $j = N$, and hence

$$C(i, k) - C(0,0) \leq C_N. \tag{9}$$

From Lemma 1, if F_j is nondecreasing,

$$p_{0j}^{k+1} F_j = p'_{0l}(p_{lj}^k F_j) \geq p_{0j}^k F_j.$$

Hence

$$C^1(0, k) = \text{Min}\left(\sum_{j=0}^{N} p_{0j}^k C_j, \sum_{j=0}^{N} p_{0j}^k L_j\right)$$

ıondecreasing in k.

ıce, by induction and (3),

$$C^n(0,k) = \text{Min}\left\{\sum_{j=0}^{N} p_{0j}^k C_j + \alpha C^{n-1}(0,0),\right.$$

$$\sum_{j=0}^{N} p_{0j}^k L_j + M + \alpha \sum_{j=0}^{N} p_{0j}^{k+1} C^{n-1}(j,0),$$

$$\left.\sum_{j=0}^{N} p_{0j}^k L_j + \alpha C^{n-1}(0,k+1)\right\}$$

ondecreasing in k.

taking limits

$$C(0,k) \text{ is nondecreasing in } k.$$

s

$$C(i,k) \geqslant C(0,k) \text{ by (4)}$$
$$\geqslant C(0,0) \text{ by (10).}$$

ıbining this with (9) we have

$$|C(i,k) - C(0,0)| \leqslant C_N.$$

hypothesis of Theorem 1.3 in Ross [16] is hence true.

We now examine (7) and (8). Note that (7) and (8) imply for some α_0 for all $\alpha_0 \leqslant \alpha < 1$,

$$F_* < 0, G_* < 0.$$

ıy Theorem 1, monotonic policies are optimal for the infinite-horizon case for all $1 > \alpha \geqslant \alpha_0$. If ıan show that all critical numbers for all $1 > \alpha \geqslant \alpha_0$ are bounded, then for $\alpha > \alpha_0$ there are a finite ber of possible optimal policies for the discounted problem. Thus we can pick a subsequence, $\to 1$, such that one monotonic policy is optimal. Combining this with Theorem 1.3 of Ross [16], see that this same policy is average-cost optimal. Furthermore, all of the critical numbers are ıded.

So we must show that critical numbers are uniformly bounded for $1 > \alpha \geqslant \alpha_0$ some α_0. By (7) and e can pick α_0 such that

$$\sup_{\alpha \geqslant \alpha_0} F_* < 0$$

and

$$\sup_{\alpha \geq \alpha_0} G_* < 0.$$

If we can show that

$$F_k = \sup_{\alpha \geq \alpha_0} F_{0k} \text{ and } G_k = \sup_{\alpha \geq \alpha_0} G_{0k}$$

converge to $\sup_{\alpha \geq \alpha_0} F_*$ and $\sup_{\alpha \geq \alpha_0} G_*$ respectively, then we can find a k_0 such that $G_{k_0} < 0$ and a k suc that $F_k < 0$ and thus $k^*(0)$ (and thus $k^*(i)$ all i) would be bounded by max $\{k_0,k\}$ for all $1 > \alpha \geq \alpha$ This would give us our result. By examining the expressions for F_{0k}, G_{0k}, F_*, and G_*, the convergenc would hold if

$$\sum_{j=0}^{N} | p_{0j}^k - \Pi_j | (C(j,0) - C(0,0)) \to 0 \tag{12}$$

and

$$C(0,k) - C_* \to 0 \tag{13}$$

uniformly in $\alpha \geq \alpha_0$. (12) holds because of relation (11). A proof of (13) is given in the Appendix.

As in the discounted case, we have sufficient conditions on the costs to assure the limiting op timality of repair (that is, (7) and (8)).

These conditions are

$$J_1 < 0 \text{ and } K_1 < 0 \tag{14}$$

where J and K have been defined for the discounted case and where we note that in computing D_{i^*} we set

$$(1-\alpha) \sum_{j=0}^{\infty} \alpha^i \sum_{l=0}^{N} p_{0l}^i L_l \Big|_{\alpha=1} = \sum_{j=0}^{N} \Pi_l L_l.$$

Since

$$\lim_{\alpha \to 1} J_\alpha = J_1 \text{ and } \lim_{\alpha \to 1} K_\alpha = K_1,$$

it is clear that from the discussion of sufficient conditions in the discounted case that (14) impli (7) and (8).

APPLICATIONS AND DISCUSSION

The results in sections 2 and 3 are useful in the general case for deteriorating systems in that the mptions and conditions involving J_α and K_α can be tested to determine whether a monotonic policy ptimal under either of the two optimality criteria. There are, however, two subclasses of deteriorating ems that are of particular interest for application.

The first class consists of all IFR and upper triangular ($p_{ij} = 0, j < i$) matrices. Note that such rices satisfy Assumption 4. It turns out in this case that the conditions $J_\alpha \leq 0$, $k_\alpha \leq 0$ together the conditions for trivial (perpetual no action) optimal policies are collectively exhaustive. That monotonic policy of some type, either trivial or nontrivial is always optimal. (Matrices of this s that satisfy an additional condition known as total positivity of order two are assumed in [15].) upper triangular assumption is of practical interest in that it stipulates that the process cannot go better (lower state) by itself, and this is often the case for a deteriorating machine or system.

The second class consists of 2×2 IFR matrices. For this class, a process is in either a good or a state and can be characterized by the following transition matrix:

$$\begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix},$$

re a is the probability of going from the good to the bad state and b is the probability of going from ad to the good state. In addition to simplicity, the class is especially interesting for two reasons. , Assumption 4 reduces to the single relation

$$a + b \leq 1.$$

nd, we can completely characterize the optimal policy. It should be noted that this class is not red by the assumptions of [15] unless $b = 0$. Ross [17] derived the optimal policy structure for problem with minor differences in convention, but only for $b=0$. This paper consequently charac- s optimal policy when $b > 0$, that is when the process can go back and forth from the good state bad state.

The complete characterization of optimal policy when $J_\alpha \leq 0$ and $k_\alpha \leq 0$ (or $J_1 < 0$ and $k_1 < 0$ ie average-cost case) consists of repair in state $(1, k)$ any k, or in state $(0, k)$, $k \geq k^*(0)$, which monotonic part. Furthermore, the region $(0, k)$, $k < k^*(0)$, can be broken up into, at most, three ional regions: a region of no action-optimal states for $0 \leq k < k_1^*$, a region of inspection for $k_1^* \leq k$ and another region of no action for $k_2^* < k < k^*(0)$ for some choice of $k_1^* \leq k_2^* \leq k^*(0)$. The is diagrammed in Figure 2. The result is summarized, along with more specific conditions in- g J_α and K_α, in:

THEOREM 3: Assume without loss of generality that $L_0 = 0$ for the 2×2 case ($N = 1$).

$$S_\alpha = C_0 b + C_1 a - a L_1 (1 + (a+b)\alpha)/(1 - \alpha + (a+b)\alpha)$$

$$T_\alpha = -M + [C_0 b + C_1 a(1-\hat{\alpha}) - L_1 \alpha + \alpha^2 a^2 L_1/(1 - \alpha + (a+b)\alpha)]/(a+b).$$

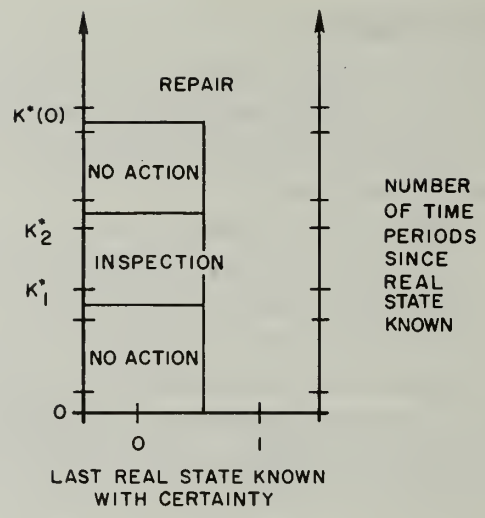


FIGURE 2. Example showing form of optimal policy for case of two real states.

Assume for the discounted criterion that $S_\alpha \leq 0$ and $T_\alpha \leq 0$, and for the average-cost criterion tha $S_1 < 0$ and $T_1 < 0$. Then under the respective criterion, when

$$a+b \leq 1,$$

a monotonic policy is optimal, and furthermore, repair is optimal in state $(1, k)$, all k, and the se $\{(0, k)\}$ can be broken up into, at most, four regions: a no-action region, an inspection region, anothe no-action region, and a repair region.

PROOF: By using the fact that

$$p_{01}^{i} = a(1-(1-a-b)^{i})/(a+b),$$

we see that by substitution,

$$S_\alpha = (a+b)J_\alpha \qquad \text{and } T_\alpha \geq K_\alpha.$$

Hence, by Theorems 1 and 2, a monotonic policy is optimal and the region $\{(1, k)\}$, $k \geq 0$ is a repair region. It remains to show that the region $\{(0, k)\}$, $0 \leq k < k^*(0)$ can be broken up into, at mos three additional regions. For the discounted criterion over a finite horizon, let

$$j_{ik}^{n} = \begin{array}{l}\text{cost of choosing inspection-cost of choosing no action} \\ \text{when an optimal policy is subsequently followed.}\end{array}$$

From the recursive relationship (2),

$$J_{0k}^n = \alpha\left(\sum_{j=0}^{1} p_{0j}^{k+1} C^{n-1}(j, 0) - C^{n-1}(0, k+1)\right) + M$$

$$= M + \alpha \underset{m \in [1, \overline{m}(n-1)]}{\text{Max}} \left\{\sum_{j=0}^{1} p_{oj}^{k+1} [C^{n-1}(j, 0) - B_{mj}^{n-1}]\right\} \text{ by (1).}$$

the function in brackets is either nonincreasing or nondecreasing in j, since $j=0$ or 1. Now if s nondecreasing in j, then

$$\sum_{j=0}^{1} p_{0j}^{k+1} F_j = \sum_{j=0}^{1} p_{0j} \sum_{l=0}^{1} p_{jl}^{k} F_l \geq \sum_{l=0}^{1} p_{0l}^{k} F_l \text{ by Lemma 1.}$$

s the function in braces is either nonincreasing or nondecreasing in k. The maximum is conse- ntly unimodal with one minimum. Since J_{0k}^n is unimodal with one minimum, $\lim_{n\to\infty} J_{0k}^n$ is unimodal one minimum. But this quantity is the cost of choosing inspection minus the cost of choosing no on (and then behaving optimally) for the discounted criterion over an infinite horizon. The assertion follows for the discounted case, since the optimal action in $(0, k)$, $0 \leq k < k^*(0)$, depends on the of $\lim J_{0k}^n$. The assertion for the average-cost criterion follows by the proof of Theorem 2. □

Two comments are relevant here. Unless C_0 and C_1 are somewhat larger than L_1, the cost condi- s will often be satisfied. The existence of "four regions" for the set $\{(0, k) : k \geq 0\}$ is a typical nt for these types of problems, and this is the subject of [15] and an important facet of [17]. As onstrated by examples in [17] and [20], four regions can certainly occur. The important result is , at most, four regions can occur.

APPENDIX

OOF THAT $C(0, k) \to C_*$ UNIFORMLY IN α

For any given $1 > \alpha \geq \alpha_0$, for k big enough, since $F_* < 0$ and $G_* < 0$,

$$C(0, k) = \sum_{j=0}^{N} p_{0j}^{k} C_j + \alpha C(0, 0) \leq \sum_{j=0}^{N} p_{0j}^{k} L_j + \alpha C(0, k+1).$$

s for $\alpha \geq \alpha_0$

$$C_* = \lim_{k\to\infty} C(0, k) = \sum_{j=0}^{N} \Pi_j C_j + \alpha C(0, 0) \leq \sum_{j=0}^{N} \Pi_j L_j + \alpha C_* \text{ from above.}$$

sing this relationship recursively,

$$C_* = \sum_{j=0}^{N} \Pi_j C_j + \alpha C(0, 0)$$

$$= \text{Min}\left\{\sum_{j=0}^{N} \Pi_j C_j + \alpha C(0, 0), \sum_{j=0}^{N} \Pi_j L_j + \alpha \sum_{j=0}^{N} \Pi_j C_j \right.$$

$$+ \alpha^2 C(0, 0), \ldots, (1-\alpha^{n+1})/(1-\alpha) \sum_{j=0}^{N} \Pi_j L_j + \alpha^{n+1} \sum_{j=0}^{N} \Pi_j C_j$$

$$\left. + \alpha^{n+2} C(0, 0), \ldots\right\}.$$

Now let $k \geq k_0$, $\alpha \geq \alpha_0$. Recall that k_0 is defined in the proof of Theorem 2 and that $G_{k_0} < 0$ for $\alpha \geq \alpha_0$. Since inspection cannot be optimal in $(0, k)$ by (12),

$$\begin{aligned} C(0,k) &= \operatorname{Min}\left\{\sum_{j=0}^{N} p_{0j}^{k} C_j + \alpha C(0,0), \sum_{j=0}^{N} p_{0j}^{k} L_j + C(0,k+1)\right\} \\ &= \operatorname{Min}\left\{\sum_{j=0}^{N} p_{0j}^{k} C_j + \alpha C(0,0), \sum_{j=0}^{N} p_{0j}^{k} L_j + \alpha \sum_{j=0}^{N} p_{0j}^{k+1} C_j \right. \\ &\quad + \alpha^2 C(0,0), \ldots, \sum_{i=0}^{n} \alpha^i \sum_{j=0}^{N} p_{0j}^{k+i} L_j + \sum_{j=0}^{N} \alpha^{n+1} p_{0j}^{k+n+1} C_j \\ &\quad \left. + \alpha^{n+2} C(0,0), \ldots\right\} \end{aligned}$$

by recursively substituting for $C(0, k+1)$, $C(0, k+2)$ and so on.

In the above minimization, n denotes how many periods to wait before repair. Since $F_* < 0$ and $G_* < 0$, the minimization will occur for some n for each α.

Since
$$|\operatorname{Min}_i (A_i) - \operatorname{Min}_i (B_i)| \leq \sup_i \{|A_i - B_i|\}$$

when the minima are attained,

$$\begin{aligned} (15) \qquad |C(0,k) - C_*| &\leq \sup\left\{\left|\sum_{j=0}^{N} (p_{0j}^{k} - \Pi_j) C_j\right|, \left|\sum_{j=0}^{N} (p_{0j}^{k} - \Pi_j) L_j \right.\right. \\ &\quad + \alpha \sum_{j=0}^{N} (p_{oj}^{k+1} - \Pi_j) C_j \Big|, \ldots, \Big|\sum_{i=0}^{m} \alpha^i \sum_{j=0}^{N} (p_{oj}^{k+i} - \Pi_j) L_j \\ &\quad \left.\left. + \alpha^{m+1} \sum_{j=0}^{N} (p_{oj}^{k+m+1} - \Pi_j) C_j \right|, \ldots\right\} \\ &\leq \sum_{i=0}^{\infty} \left\{\left|\sum_{j=0}^{N} p_{oj}^{k+i} - \Pi_j\right| L_j + \sup_{m \geq 0} \sum_{j=0}^{N} \left|p_{oj}^{k+m} - \Pi_j\right| C_j. \right. \end{aligned}$$

Recalling Assumption 3 and the ergodic theorem (e.g., Fisz [6]),

$$|p_{0j}^{k+i} - \Pi_j| \leq C \cdot (1-\delta)^{(k+i)/r} \text{ where } C \text{ is a constant.}$$

By substituting into (15),

$$|C(0,k) - C_*| \leq \text{constant} \times (1-\delta)^k \to 0 \text{ uniformly.}$$

ACKNOWLEDGEMENT

The author would like to thank G. J. Lieberman of Stanford University for supervising this researc while the author was a graduate student and D. A. Butler and others at Stanford University and th State University of New York at Stony Brook for helping review the manuscript.

BIBLIOGRAPHY

Blackwell, D., "Discounted Dynamic Programming," Annals of Mathematical Statistics *36*, 226–235 (1965).

Derman, C., "On Optimal Replacement Rules When Changes of State are Markovian," *Mathematical Optimization Techniques*, R. Bellman, ed. (University of California Press, Berkeley, 1963).

Derman, C., "Markovian Sequential Control Processes—Denumerable State Space," Journal of Mathematical Analysis and Applications *10*, 295–302 (1965).

Derman, C., *Finite State Markovian Decision Processes* (Academic Press, New York, 1970).

Emmons, H. and J. Poimboeuf, "Optimal Markovian Replacement with Inspection Costs," Technical Report No. 131, Department of Operations Research, Cornell University, Ithaca, N.Y. (1971).

Fisz, M., *Probability Theory and Mathematical Statistics* (John Wiley and Sons, Inc., New York, 1963).

Girshick, M. A. and H. Rubin, "A Bayes Approach to a Quality Control Model," Annals of Mathematical Statistics *23*, 114–125 (1952).

Karlin, S., *Total Positivity* (Stanford University Press, Stanford, Calif., 1968), Vol. 1.

Karlin, S., *A First Course in Stochastic Processes* (Academic Press, New York, 1969).

Klein, M., "Inspection-Maintenance-Replacement Schedules under Markovian Deterioration," Management Science *9*, 25–32 (1962).

Kolesar, P., "Minimum Cost Replacement Under Markovian Deterioration," Management Science *12*, 694–706 (1966).

Lehmann, E., Testing Statistical Hypotheses (John Wiley and Sons, Inc., New York, 1960).

Ray, D. K., "Replacement Rules for Partially Observable Equipment," Technical Report No. 37, Operations Research Center, M.I.T., Cambridge, Mass. (1967).

Rosenfield, D. B., "Deteriorating Markov Processes Under Uncertainty," Technical Report No. 162, Department of Operations Research and Department of Statistics, Stanford University, Stanford, Calif. (1974).

Rosenfield, D. B., "Markovian Deterioration with Uncertain Information," Operations Research *24*, 141–155 (Jan.–Feb. 1976).

Ross, S. M., "Non-Discounted Denumberable Markovian Decision Models," Annals of Mathematical Statistics *39*, 412–423 (1968).

Ross, S. M., "Quality Control Under Markovian Deterioration," Management Science *17*, 694–706 (1971).

Smallwood, R. D. and E. J. Sondik, "The Optimal Control of Partially Observable Markov Processes Over a Finite Horizon," Operations Research *21*, 1071–1088 (1973).

Tafeen, S., "A Markov Decision Process with Costly Inspection," Technical Report No. 107, Department of Operations Research and Department of Statistics, Stanford University, Stanford, Calif. (1968).

Taylor, H. M., "Markovian Sequential Replacement Processes," Annals of Mathematical Statistics *36*, 1677–1694 (1965).

AN ALGORITHM FOR THE SEGREGATED STORAGE PROBLEM

A. W. Neebe
University of North Carolina at Chapel Hill
and
M. R. Rao
University of Rochester
Rochester, New York

ABSTRACT

The segregated storage problem involves the optimal distribution of products among compartments with the restriction that only one product may be stored in each compartment. The storage capacity of each compartment, the storage demand for each product, and the linear cost of storing one unit of a product in a given compartment are specified. The problem is reformulated as a large set-packing problem, and a column generation scheme is devised to solve the associated linear programming problem. In case of fractional solutions, a branch and bound procedure is utilized. Computational results are presented.

NTRODUCTION

Consider a set of products indexed by $I=\{1, 2, \ldots m\}$, each with known storage demand $a_i > 0$, set of compartments indexed by $J=\{1, 2, \ldots, n\}$, each with known storage capacity $b_j > 0$. Segregated Storage Problem (SSP) is to minimize total storage costs subject to the conditions that oducts be stored, and that at most one product be stored in any compartment. One example SSP is the silo problem, where a number of different grains are to be stored in the compartments silo. Clearly, no compartment may simultaneously contain more than one variety of grain. Ex-s of the SSP are given by Eilon and Mallya [2], Shlifer and Naor [9], White [10], and White and is [11].

ince all demand must be met, it is assumed that external storage space (an $(n+1)^{st}$ compartment) ilable at a premium. Defining $J'=J \cup (n+1)$, note that more than one product may be stored ually. Let c_{ij}, $i \in I$, $j \in J'$, be the per unit storage cost of product i in compartment j. Let x_{ij}, $i \in I$, be the number of units of product i to be stored in compartment j.

hen the SSP is to

$$\text{minimize} \quad \sum_{i \in I} \sum_{j \in J'} c_{ij} x_{ij},$$

t to

$$\sum_{j \in J'} x_{ij} = a_i \qquad \text{for all } i \in I$$

$$\sum_{i \in I} x_{ij} \leq b_j \qquad \text{for all } j \in J \tag{3}$$

$$x_{ij} \cdot x_{kj} = 0 \qquad \text{for all } i \in I,\ k \in I,\ i \neq k \text{ and all } j \in J \tag{4}$$

$$x_{ij} \geq 0 \qquad \text{for all } i \in I,\ j \in J'. \tag{5}$$

Constraints (2) insure that all products are stored. Constraints (3) insure that storage capacity is n exceeded. Constraints (4) guarantee that at most one product is stored in a compartment. Observ that since all of the m products can be allocated to external storage, the *SSP* always contains at lea one feasible solution.

Notice that (1)–(3), (5) is a transportation problem. White and Francis [11] take advantage of th fact and propose a branch and bound algorithm which solves a transportation problem at every noc of the branch and bound tree. For an alternate approach, they suggest an extreme point ranking proc dure. No computational results are given for either algorithm. However, a wealth of computation experience on a branch and bound algorithm is provided by Dannenbring and Khumawala [1]. The compare various combinations of node selection rules and branching decisions rules, and furthermor give computational results for a number of heuristic techniques. An implicit enumeration algorith which takes advantage of the graph theoretic properties of the *SSP* is developed by Evans [3]. A mo erate amount of computational results is recorded.

In the following sections, we present a different formulation of the SSP, together with a propose algorithm for its solution, an example, and computational results. The computational results indica that our algorithm is substantially faster than Dannenbring and Khumawala's algorithm.

2. AN ALTERNATE FORMULATION

Define an *allocation* of product i to be any allocation of all a_i units of product i among the $n+1$ compartments such that the capacity of no compartment is exceeded. Thus in general any allocati can be represented by the vector $w_i = (w_{i1}, w_{i2}, \ldots, w_{i,n+1})$ where in addition

$$\sum_{j \in J'} w_{ij} = a_i$$

and

$$0 \leq w_{ij} \leq b_j \qquad \text{for all } j \in J'.$$

The cost of allocation w_i equals $\sum_{j \in J'} c_{ij} w_{ij}$.

Any allocation w_i does not make efficient use of the compartments if the lesser-cost compartmen used by w_i are not filled to capacity. For this reason we can restrict our attention to the so-called *ef cient allocations*. Intuitively, the efficient allocations dominate (in either a cost sense or a compartme utilization sense) all nonefficient allocations. An efficient allocation is defined as follows: Let $S(w_i)$

$_{ij} > 0\}$ be the set of all compartments which are utilized by product i using allocation w_i. Then w_i id to be *efficient* if

or all $j \epsilon J'$, $w_{ij} = 0$ or b_j, except for at most one $j = j^*$
here $0 < w_{ij^*} < b_{j^*}$, and
or all $j \epsilon J'$, if $w_{ij} > 0$, then $w_{ih} = b_h$ for all $h \epsilon S(w_i)$,
uch that $c_{ih} < c_{ij}$.

he remainder of this paper we will let $y_i^k = (y_{i1}^k, y_{i2}^k, \ldots, y_{i,n+1}^k)$ denote the kth efficient allocation oduct i.

An efficient allocation y_i^k can be obtained from (nonefficient) allocation w_i in the following fashion: the costs c_{ij}, $j \epsilon S(w_i)$, be ranked in nondecreasing order $c_{i(1)} \leq c_{i(2)} \leq \ldots$, where ties may be en arbitrarily. Clearly, the least cost way of allocating product i among the compartments in) is to allocate $y_{i(1)}^k = b_{(1)}$ units to compartment (1), $y_{i(2)}^k = b_{(2)}$ units to compartment (2), and so on, all a_i units are allocated. Assume compartment (l) is the last compartment to be allocated any oduct i. Thus (l) is the smallest index h, such that $\sum_{j=1}^{h} b_{(j)} \geq a_i$. Then compartment (l) is allocated $= a_i - \sum_{j=1}^{l-1} b_{(j)}$ units, while $y_{i(j)}^k = b_{(j)}$ for all $j < l$, and $y_{i(j)}^k = 0$ for all $j > l$. The cost of y_i^k equals

$$c_i^k = \sum_{j=1}^{l-1} c_{i(j)} b_{(j)} + c_{i(l)} \left[a_i - \sum_{j=1}^{l-1} b_{(j)} \right].$$

ly, the cost of y_i^k is no greater than the cost of w_i.

Assume there are n_i efficient allocations of product i, and let $N_i = \{1, 2, \ldots, n_i\}$. Define decision bles z_i^k, $i \epsilon I$, $k \epsilon N_i$, such that

$$z_i^k = \begin{cases} 1 \text{ if efficient allocation } y_i^k \text{ is used} \\ 0 \text{ if otherwise} \end{cases}$$

ll $i \epsilon I$, $j \epsilon J$, and $k \epsilon N_i$ define a vector $g_i^k = (g_{i1}^k, g_{i2}^k, \ldots, g_{in}^k)$, such that

$$g_{ij}^k = \begin{cases} 1 \text{ if } y_{ij}^k > 0 \\ 0 \text{ otherwise} \end{cases}.$$

the ones in the vector g_i^k identify the compartments that are assigned to product i using efficient ation y_i^k. Observe that while the vector y_i^k is of length $(n+1)$, the corresponding vector g_i^k is only ngth n. This is the case because, in the SSP, it is irrelevant how many different products are allo- to the external compartment.

The new formulation of the SSP is to

inimize $\sum_{i \epsilon I} \sum_{k \epsilon N_i} c_i^k z_i^k,$

subject to

$$\sum_{i \in I} \sum_{k \in N_i} g_{ij}^k z_i^k \leq 1 \qquad \text{for all } j \in J \tag{8}$$

$$\sum_{k \in N_i} z_i^k = 1 \qquad \text{for all } i \in I \tag{9}$$

$$z_i^k = 0 \text{ or } 1 \qquad \text{for all } i \in I,\ k \in N_i. \tag{10}$$

The constraints (8) insure that no compartment is assigned more than one product, while the constraints (9) guarantee that every product is fully allocated. Observe that (7)-(10) is a set-packing problem with the additional constraint (9) specifying that exactly one efficient allocation for every product is used. If the binary constraints (10) are relaxed to

$$z_i^k \geq 0, \tag{11}$$

a linear program (LP) is obtained. For the remainder of this paper we will refer to (7)–(9), (11) as the LP version of the SSP. We will show how this LP may be solved using a column generation technique in which only the "best" (with respect to cost) columns are generated.

3. SOLUTION OF THE LINEAR PROGRAM

Let B be any feasible basis to (8), (9), (11). The matrix B contains $(n+m)$ rows. An additional "objective function" row is included to obtain the Simplex multipliers. Any column of B corresponding to efficient allocation y_i^k is of the form $(g_{i1}^k, g_{i2}^k, \ldots, g_{in}^k, e_i)^\tau$, where e_i is the ith unit vector and τ denotes transpose.

Any feasible solution to the SSP can be converted into an easily invertible, initial, feasible basis B_0. Let $(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m)$ be the efficient allocations of the m products associated with any feasible solution to the SSP. A feasible solution may be obtained by executing any heuristic of the SSP. Let $\hat{g}_i = (\hat{g}_{i1}, \hat{g}_{i2}, \ldots \hat{g}_{in})$, as defined in (6), correspond to $\hat{y}_i$, $i \in I$. Let the ith column of B_0 equal $(\hat{g}_i, e_i)^\tau$, $i \in I$. Let the $(m+j)$th column of B_0 equal $(e_j, \phi)^\tau$, $j \in J$, corresponding to the jth slack variable of (8), where ϕ is a vector containing m zeroes. Thus if

$$\hat{G} = [\, \hat{g}_1^\tau, \hat{g}_2^\tau, \ldots, \hat{g}_m^\tau \,],$$

then

$$B_0 = \begin{bmatrix} \hat{G} & I \\ I & 0 \end{bmatrix}, \tag{12}$$

where I is the identity matrix of appropriate size, and 0 is a matrix of all zeroes. The inverse is readily seen to be

$$B_0^{-1} = \begin{bmatrix} 0 & I \\ I & -\hat{G} \end{bmatrix}.$$

The LP version of the SSP would be difficult to solve if it were necessary to generate all of the ient allocations and thus all of the columns in the LP. Fortunately, a column generation procedure be employed to solve the LP. In this way, the LP columns are generated only as they are needed. s procedure is used in conjunction with the Revised Simplex algorithm, which maintains a working is inverse of order $(m+n+1)$. See Lasdon [7] for a more complete discussion of column generation niques. For an example of column generation applied to the cutting-stock problem, see Gilmore Gomory [5]. Our notion of an efficient allocation is similar to their concept of a cutting activity.

The vector $(\sigma,\pi) = (\sigma_1, \ldots, \sigma_n, \pi_1, \ldots, \pi_m) = C_B B^{-1}$ is found in the objective function row of current Revised Simplex tableau. The vector σ contains the Simplex multipliers corresponding 3), while the vector π contains the multipliers corresponding to (9). The vector C_B gives the appro- te costs of the columns in the current basis.

The basis corresponding to any LP tableau is optimal if all nonbasic variables have nonpositive uced cost. Thus (7)–(9), (11) has been solved if, for all nonbasic columns $(g_i^k, e_i)^\tau$, $i \epsilon I$, $k \epsilon N_i$, $\tau)\ (g_i^k, e_i)^\tau - c_i^k \leq 0$. Otherwise bring into the basis nonbasic column $(g_p^q, e_p)^\tau$, such that

$$(\sigma,\pi)\ (g_p^q, e_p)^\tau - c_p^q = \max_{\substack{i \epsilon I \\ k \epsilon N_i}} \{(\sigma,\pi)\ (g_i^k, e_i)^\tau - c_i^k\} > 0.$$

The column $(g_p^q, e_p)^\tau$ represents the nonbasic column having the largest positive reduced cost, and btained by solving m subproblems (corresponding to the m products). An efficient dynamic program- g algorithm for solving these subproblems is given in a following section. In solving the ith subprob- , we either obtain the nonbasic column (corresponding to product i) having the largest positive re- ed cost, or else determine that no nonbasic column (corresponding to product i) has a positive re- ed cost.

The subproblem for product i, $i \epsilon I$, is to

$$\text{maximize } \theta_i = \sum_{j \epsilon J'} f_{ij}(t_{ij}) + \pi_i,$$

ect to

$$\sum_{j \epsilon J'} t_{ij} = a_i$$

$$0 \leq t_{ij} \leq b_j \qquad \text{for all } j \epsilon J$$

$$0 \leq t_{i,n+1}$$

e for $j \epsilon J'$

$$f_{ij}(t_{ij}) = \begin{cases} 0 \text{ if } t_{ij}=0 \\ \sigma_j - c_{ij}t_{ij} \quad \text{if } t_{ij}>0 \end{cases} \tag{18}$$

and where $\sigma_{n+1}=0$.

If $t_i^*=(t_{i1}^*, t_{i2}^*, \ldots, t^*_{i,n+1})$ is an optimal solution to (14)–(17) with value θ_i^*, determine p such th $\theta_p^*=\max_{i\in I}\theta_i^*$, with ties broken arbitrarily. If $\theta_p^*\leq 0$, the LP (7)–(9), (11) is optimal. If $\theta_p^*>0$, the colun introduced into the basic is $(g_p, e_p)^\tau$ where $g_p=(g_{p1}, g_{p2}, \ldots, g_{pn})$ is such that

$$g_{pj}=\begin{cases} 1 \text{ if } t_{pj}^*>0 \\ 0 \text{ otherwise} \end{cases}.$$

One well-known result of mathematical programming is that the maximum of a convex functi subject to a set of linear inequalities occurs at an extreme point of the feasible region. (See Hadley for discussion of this result.) All of the subproblems possess convex objectives functions (14) if $\sigma_j \leq$ for all $j\in J$. If there exists some $\sigma_p>0$, the pth slack variable corresponding to (8) has a positive reduc cost and may be introduced into the basis. The column introduced would be $(e_p, \phi)^\tau$. This process c be repeated until the desired result is obtained. The lexicographic ordering rule can be used in t event of degeneracy to avoid cycling, although we have experienced no problem with cycling in the ma problems solved. (Refer to Garfinkel and Nemhauser [4] for a discussion of cycling and the lexicograph ordering rule.)

An extreme point of the feasible region (15)–(17) is such that, at most, one $0<t_{ij}<b_j$. Converse every solution to (15)–(17) such that, at most, one $0<t_{ij}<b_j$ is an extreme point. Thus every efficie allocation corresponding to product i is an extreme point of (15)–(17). However, (15)–(17) may ha extreme points which are nonefficient allocations. But an efficient allocation always maximizes (1 subject to (15)–(17), since each nonefficient allocation has an objective function value no great than the corresponding efficient allocation. Further, the reduced cost value for the kth efficient alloc tion is given by $(\sigma, \pi)\ (g_i^k, e_i)^\tau - c_i^k$. Thus maximizing θ_i is equivalent to maximize $\{(\sigma, \pi)\ (g_i^k, e_i)^\tau - c_i^k$ $k\in N_i$

Observe that in solving (14)–(17) for all $i\in I$, we are solving for an efficient allocation t_p^* having tot cost equal to

$$c_p^*=\sum_{j\in J'} c_{pj}t_{pj}^*.$$

However, the corresponding linear programming column $(g_p, e_p)^\tau$ has total reduced cost equal to

$$\left[\sum_{j\in J'}\sigma_j g_{pj}\right]+\pi_p-c_p^*.$$

This explains the "fixed charge" shape of the objective function (14).

The column generation algorithm to solve (7)–(9), (11) is given below.

ALGORITHM:

(Step 1) Starting with any feasible solution of the SSP, obtain the initial basis inverse B_0^{-1} and t initial vector of Simplex multipliers (σ, π). Go to step 2.

(Step 2) If $\sigma_j \leq 0$ for all $j \epsilon J$, go to step 3. Otherwise, let p be such that $\sigma_p = \max_{j \epsilon J} \{\sigma_j\}$. The column $\phi)^T$ should be introduced into the basis. Go to step 4.

(Step 3) Examine the subproblems (14)–(17) for all $i \epsilon I$ to try to find a column with positive reduced . If no column with positive reduced cost exists, stop (the optimal solution to (7) –(9), (11) has been ined). Otherwise, $(g_p, e_p)^T$, a generated column having positive reduced cost, should be introduced the basis, and go to step 4.

(Step 4) Perform the pivot operation, obtaining a new basis inverse and a new vector of Simplex ipliers. Drop the column which leaves the basis, and go to step 2.

EXAMPLE

Consider an example problem similar to that given by White and Francis [11] and also by Dannen- g and Khumawala [1].

i \ j	compartment 1	2	3	4	demand a_i
1	20	14	19	24	1
product 2	15	13	20	22	8
3	18	18	15	22	7
capacity b_j	3	7	4	16	

The only difference from their example is that we have expressed all demands a_i and capacities terms of multiples of 10 units. The fourth compartment represents external storage. The efficient ations for product i, $i \epsilon I$, were generated by considering all possible subsets of J' having total city greater than, or equal to, the demand a_i. For each of these subsets, an efficient allocation obtained by utilizing only the least costly of the selected compartments. This is done only for illus- ve purposes, since the proposed algorithm generates columns only as they are needed. The 18 rent efficient allocations, along with their respective costs, are enumerated columnwise in Table

TABLE 1.

	Product $i=1$				Product $i=2$							Product $i=3$						
	k				k							k						
	1	2	3	4	1	2	3	4	5	6	7	1	2	3	4	5	6	7
$(y_i^k)^T$:	1	0	0	0	0	0	0	0	3	1	3	0	0	0	0	3	3	3
	0	1	0	0	0	0	7	7	0	7	0	0	7	0	3	0	0	4
	0	0	1	0	0	4	0	1	0	0	4	0	0	4	4	0	4	0
	0	0	0	1	8	4	1	0	5	0	1	7	0	3	0	4	0	0
c_i^k:	20	14	19	24	176	168	113	111	155	106	147	154	126	126	114	142	114	126

Assume that a solution to the SSP is obtained using some heuristic. In this solution the 1 unit product 1 is allocated to external storage, while 1 and 7 units of product 2 are allocated to compartmer 1 and 2, respectively. Finally, 4 units of product 3 are allocated to compartment 3, and the remaini 3 units are allocated to external storage. This solution corresponds to efficient allocations y_1^4, y_2^6, and y and has a total cost of 256. The initial basis corresponding to (12) is

$$B_0 = \left[\begin{array}{ccc:ccc} 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ \hdashline 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{array}\right].$$

Three subproblems of the form (14)–(17) must now be solved to determine that column, if ar having largest positive reduced cost. For subproblem 1 (product 1), $t_1^* = y_1^2$ solves (14)–(17), and h reduced cost equal to 10. For subproblem 2, $t_2^* = \emptyset$, as no column has positive reduced cost. For su problem 3, $t_3^* = y_3^4$ solves (14)–(17), and has reduced cost equal to 12. Hence the column introduc into the basis B_0 is $(g_3^*, e_3)^T$ where g_3^* corresponds to t_3^*.

After four additional columns are generated, and hence after four additional pivot operatio the optimal linear programming solution is obtained: $z_1^4 = z_2^3 = z_3^6 = 1$ having an objective functi value of 251. Since the solution is all-integer, this indicates that the optimal procedure is to utili efficient allocations y_1^4, y_2^3, *and* y_3^6.

5. SOLUTION OF THE SUBPROBLEMS

As stated, the subproblems (14)–(17) can be solved using dynamic programming. However, is possible to devise a particularly efficient dynamic programming recursion by taking advantage the fact that some efficient allocation solves (14)–(17). Specifically, we take advantage of the fact th for any efficient allocation y_i^k, if $0 < y_{ij^*}^k < b_{j^*}$, then $y_{ij}^k = 0$ for all j such that $c_{ij} > c_{ij^*}$.

Consider the following recursion for any subproblem (product) i. Let the costs c_{ij}, $j \epsilon J'$, be rank in nondecreasing order $c_{i(1)} \leqslant c_{i(2)} \leqslant \ldots \leqslant c_{i,(n+1)}$. In general, $F_j(k)$ is the maximum (partial) r duced cost of being in state k with stages $j+1, j+2, \ldots, n, n+1$ yet to be considered. The sta variable k identifies the number of units of product i that have been allocated in stages $j, j-1, \ldots,$ The stages, of course, correspond to the compartments. The calculations start at stage 1 and proce to stage $n+1$. The maximum reduced cost of all efficient allocations is given by $F_{n+1}(a_i)$.

culations for stage $j = 1$

$F_1(0) = \pi_i$, $F_1(k) = -\infty$ for $k = 1,2, \ldots, a_i$.

If $b_{(1)} < a_i$, let $F_1(b_{(1)}) = \sigma_{(1)} - c_{i,(1)} \cdot b_{(1)} + \pi_i$.

Otherwise, let $F_1(a_i) = \sigma_{(1)} - c_{i,(1)} \cdot a_i + \pi_i$.

culations for stage $j = 2, 3, \ldots, n+1$

$F_j(k) = F_{j-1}(k)$ for $k = 0, 1, \ldots, a_i$.

If $b_{(j)} \geq a_i$, let $l = 0$ and go to step 2.

Otherwise go to step 1.

p 1)

$$F_j(k + b_{(j)}) = \max [F_j(k + b_{(j)}); \\ F_{j-1}(k) + \sigma_{(j)} - c_{i,(j)} \cdot b_{(j)}] \\ \text{for } k = 0, 1, 2, \ldots, a_i - b_{(j)} - 1.$$

$l = a_i - b_{(j)}$ and go to step 2.

p 2)

$$F_j(a_i) = \max [F_j(a_i); \\ \max_{l \leq k \leq a_i - 1} (F_{j-1}(k) + \sigma_{(j)} - c_{i,(j)} \cdot (a_i - k))].$$

(End of calculations for stage j).

A brief explanation of the recursion is helpful. If at stage j we are in state $k < a_i$, then $t_{i,(j)}$ can equal 0 or $b_{(j)}$. These calculations are carried out in step 1. On the other hand, if at stage j we n state $k = a_i$, then $t_{i,(j)}$ can assume values between 0 and $b_{(j)}$ inclusive. These calculations are ed out in step 2. If $F_{n+1}(a_i) > 0$, an efficient allocation with positive reduced cost has been found. Instead of solving (14)–(17) exactly, a good heuristic solution can be easily obtained in the following ner. For any product i, rank the compartments j in nonincreasing order $[\sigma_{(1)} - c_{i(1)}b_{(1)}]/b_{(1)} \geq - c_{i(2)}]/b_{(2)} \geq \ldots$. Ties may be broken arbitrarily. Then allocate $b_{(1)}$ units to compartment $_{(2)}$ units to compartment (2), and so on until all demand a_i is satisfied. The last compartment to ssigned, say compartment (l), will be allocated an amount less than, or equal to, $b_{(l)}$ units. Ob- that if compartment (l) is allocated exactly $b_{(l)}$ units, then the heuristic optimally solves (14)–(17). This is easily seen. Suppose the objective function (14) is replaced by

maximize $\sum_{j \in J'} d_{ij} t_{ij} + \pi_i$

where $d_{ij} = (\sigma_j - c_{ij}b_j)/b_j$. Maximizing (14a) subject to (15)–(17) gives an upper bound for max θ. Clearly the heuristic procedure solves (14a), (15)–(17).

Thus $\pi_i + \sum_{j=1}^{l} (\sigma_{(j)} - c_{i(j)})\, b_{(j)}$ provides an upper bound for max θ_i. If compartment (l) is allocated exactly $b_{(l)}$ units, the heuristic obtains a feasible solution with an objective value equal to the bound. Our column generation heuristic is an example of a "greedy" algorithm. See Magazine et al. [8] for an example of another greedy algorithm.

From a computational point of view, it may be better to use the heuristic to generate columns, reserving the dynamic programming algorithm for cases in which no column with positive reduced cost can be found. The greater number of pivot operations that in general will be required may be offset by the faster column generating procedure.

6. EXAMPLE ,Continued.

Consider the SSP of the previous example. After the first pivot operation, the basis inverse is

$$B^{-1} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix}$$

The basic variables correspond to y_1^4, y_2^6, y_3^3, s_1, y_3^4 and s_3 (in that order), where s_1 and s_3 are slack variables. Thus $(\sigma, \pi) = C_B B^{-1} = (24, 106, 126, 0, 114, 0)B^{-1} = (0, -12, 0, 24, 118, 126)$. The dynamic programming recursion corresponding to product $i=3$ is as follows. First, the costs c_{3j}, $j=1, 2, 3, 4$ are ranked in nondecreasing order so that $c_{3,(1)} = c_{3,3} = 15$, $c_{3,(2)} = c_{3,1} = 18$, $c_{3,(3)} = c_{3,2} = 18$ and $c_{3,(4)} = c_{3,4} = 22$.

The final values for $F_j(k)$ for all states $k=0, \ldots, 7$ and for all stages $j=1, \ldots, 4$ are given in Table 2.

TABLE 2

j \ k	0	1	2	3	4	5	6	7
1	126	$-\infty$	$-\infty$	$-\infty$	66	$-\infty$	$-\infty$	$-\infty$
2	126	$-\infty$	$-\infty$	72	66	$-\infty$	$-\infty$	12
3	126	$-\infty$	$-\infty$	72	66	$-\infty$	$-\infty$	12
4	126	$-\infty$	$-\infty$	72	66	$-\infty$	$-\infty$	12

ce $F_4(7)=12>0$, we have found a column with a positive reduced cost. Working backwards ough the recursion, the corresponding optimal solution is $t^*_{3,4}=0$, $t^*_{3,2}=0$, $t^*_{3,1}=3$, and $t^*_{3,3}=4$, ch corresponds to efficient allocation y^6_3.

Now using the heuristic technique,

$$[\sigma_1-c_{3,1}b_1]/b_1=[0-18\cdot 3]/3=-18$$

$$[\sigma_2-c_{3,2}b_2]/b_2=[-12-18\cdot 7]/7=-19.7$$

$$[\sigma_3-c_{3,3}b_3]/b_3=[0-15\cdot 4]/4=-15$$

$$[\sigma_4-c_{3,4}b_4]/b_4=[0-22\cdot 16]/16=-22.$$

r units of product 3 are assigned to compartment 3 and the remaining 3 units to compartment 1. total reduced cost of this efficient allocation is $-60-54+\pi_3=-114+126=12$. Observe that is the same efficient allocation generated using the dynamic programming algorithm. Of course, is not always the case.

RESOLVING FRACTIONAL TERMINATION

If the LP version of the SSP terminates fractional, then at least one compartment $q\in J$ is assigned nore than one product. Let $p\in I$ be any of these products. One possible way of obtaining an integer ition is to utilize a branch and bound scheme starting with x_{pq}. Two LP subproblems of the form (9), (11) are created. In the first subproblem, compartment q is prohibited from being assigned to duct p. This may be accomplished by setting $c_{pq}=R$, where R is a sufficiently large positive cost. effect, this eliminates all efficient allocations y^k_p, $k\in N_p$, having $y^k_{pq}>0$. In the second subproblem, partment q is not assigned to any product except possibly p. This may be accomplished by setting $=R$ for all $i\in I$ except $i=p$. Observe that this is not a strict partition of the feasible region, as it is sible for compartment q not to be assigned to product p in both subproblems. However, a strict ition may be obtained if in the second subproblem we eliminate the slack variable associated with partment q.

The two LP subproblems can be solved by the column generation algorithm outlined in the pre- s sections. Hopefully both LP subproblems will terminate in all-integer solutions. However, if er terminates in a fractional solution that has an objective value less than the objective value of best currently available integer solution, then additional LP subproblems must be created and ed. The branch and bound procedure continues until a minimum cost, all-integer solution is ined.

Before solving any LP subproblems, it makes sense to check whether the fractional solution can sed to easily obtain an optimal integer solution. The following procedure, which takes advantage e fact that there usually exists at least one alternate optimal solution which is all-integer, was found e very effective. Let $\{\hat{x}^k_i\}$ be any (fractional) LP solution with objective value $\hat{z}$, and let $\{\hat{y}^k_i\}$ be the esponding set of efficient allocations. Consider in turn all $\hat{x}^k_i$ which are equal to 1. (Every LP frac- al solution we have encountered so far has contained a number of values equal to 1.) Each such

$\hat{x}_i^k$ identifies a set of compartments $S(\hat{y}_i^k)$ which, except possibly for the external storage compartmen are assigned only to product i. Remembering that efficient allocation $\hat{y}_i^k$ is to be used in storing produc i, eliminate product i and all internal compartments in $S(\hat{y}_i^k)$ from the problem. Proceeding in thi fashion, a SSP of reduced size is obtained. The same heuristic used to obtain the initial feasible solu tion to the SSP may be applied to this reduced problem. An optimal solution to the original SSP ha been obtained if $\hat{z}$ equals the heuristic cost of the reduced problem plus the cost of the previously estal lished efficient allocations.

8. COMPUTATIONAL RESULTS

The proposed SSP algorithm was written in double precision FORTRAN IV, compiled under leve H, and run on an IBM 370/165. The "demand allocation heuristic" suggested by Dannenbring an Khumawala [1] was used to obtain an initial basis inverse. Throughout the calculations the invers was stored as an adjoint matrix and a determinant, thus eliminating roundoff errors in the calculations At every LP iteration the column generation heuristic was employed to attempt to find a column wit a positive reduced cost. Only when it failed was the dynamic programming algorithm used.

If the LP version of the SSP terminated fractional, the procedure outlined in the previous sectio was used to try to easily obtain an alternate optimal integer solution. In this case, the demand allocatio heuristic was the solution heuristic applied to the reduced SSP. If the procedure failed to obtain a optimal integer solution, the appropriate modified-cost data was generated, and the necessary L branch and bound subproblems were solved from scratch using our SSP algorithm. Because of th relatively few number of LP subproblems that were required in the computations, this technique i probably more efficient than the programming of a more elaborate branch and bound procedure.

A total of 38 data sets were randomly generated, each data set being composed of a number c problems of identical size. The first 9 data sets are based on the data sets used by Dannenbring an Khumawala [1]. The only difference from their data is that we have expressed all demands a_i and capac ities b_j in terms of multiples of 10. Since their values were in whole multiples of 10, no essential change have been made. However, the effectiveness of our dynamic programming column generation procedur can be increased by using smaller values of a_i and b_j. The next 9 data sets were generated in a simila fashion using the same data generator. The parameters of the uniform distributions used in generatin the data are as follows. Let $U[p,q]$ denote the discrete uniform distribution between p and q, inclusive Then, for the first 18 data sets (called Type 1 data sets), the internal storage costs c_{ij} were drawn fron $U[10,19]$, the external storage costs $c_{i,n+1}$ were drawn from $U[20,24]$, and the capacities a_i and de mands b_j were all drawn from $U[1,9]$.

The remaining 20 data sets were constructed to test the effectiveness of our SSP algorithm i solving problems in which the data had little variation. For all of these problems, the internal storag costs were drawn from $U[13,15]$, while the external storage costs were all set equal to 20. For Type data sets, the capacities were drawn from $U[1,9]$, and the demands from $U[5,6]$. For Type 3 data sets the capacities were drawn from $U[5,6]$, and the demands from $U[1,9]$.

Computational results are given in Tables 3 to 5. For all data sets we indicate the problem size the number of problems, and the number for which an optimal integer solution was readily available This includes problems which terminated all-integer, and problems which terminated fractional bu from which an optimal integer solution was easily obtained. We indicate the average and maximun

ıber of pivot operations, and the average and maximum c.p.u. time in seconds required to solve primary LP. These figures are exclusive of input/output, and exclusive of time needed for solving LP subproblems. The next statistic gives the total number of LP subproblems which had to be ed to determine optimal integer solutions for all problems in the data set. The final statistic, when lable, is the average solution time (also in c.p.u. seconds on the IBM 370/165) required by the ıch and bound algorithm of Dannenbring and Khumawala [1].

In summary, all of the 251 problems attempted were solved. Type 2 and Type 3 problems were erally easier to solve than Type 1 problems, both in terms of the effort required to solve the primary and in the number of LP subproblems that had to be solved. Actually, Type 2 and Type 3 problems a greater tendency to terminate fractional. However, in almost all cases an optimal integer solu- could be derived easily from the fractional solution.

TABLE 3. *Type 1 Data Sets*

Data set	m	$n+1$	Total problems	Total integer	Average pivots	Maximum pivots	Average time	Maximum time	Total LP subproblems	D & K average time
1	10	10	10	All	26	35	0.17	0.13	—	
2	15	5	10	All	14	25	0.08	0.12	—	0.10
3	20	5	10	All	13	17	0.09	0.13	—	0.20
4	25	5	10	All	14	19	0.12	0.16	—	0.20
5	5	15	10	All	1	5	0.02	0.05	—	0.60
6	5	20	10	All	2	5	0.03	0.05	—	0.90
7	5	25	10	All	1	5	0.03	0.06	—	0.70
8	12	12	10	8	53	105	0.39	0.72	6	6.80*
9	15	15	10	9	77	117	0.82	1.24	6	15.20†
10	20	20	10	All	130	238	2.37	3.86	—	
11	25	25	10	All	224	344	5.96	9.50	—	
12	20	40	10	7	27	91	1.04	3.32	4	
13	40	20	10	All	164	231	5.43	6.87	—	
14	30	30	10	7	334	510	12.23	18.26	4	
15	40	40	5	4	633	867	39.97	54.99	1	
16	30	50	5	4	72	107	3.57	5.62	1	
17	50	30	5	All	371	46	20.63	24.62	—	
18	50	50	2	0	992	1124	94.98	105.95	6	

*Average time for 8 or 10 problems solved.
† Average time for 3 or 10 problems solved.

TABLE 4. *Type 2 Data Sets*

Data set	m	$n+1$	Total problems	Total integer	Average pivots	Maximum pivots	Average time	Maximum time	Total LP subproblems
19	15	15	5	4	68	73	0.89	1.06	1
20	20	20	5	4	85	98	1.76	1.89	1
21	25	25	5	All	130	167	3.94	4.80	—
22	20	40	5	All	0	0	0.20	0.21	—
23	40	20	5	All	88	111	3.17	4.29	—
24	30	30	5	All	185	211	7.77	9.33	—
25	40	40	5	All	353	453	27.02	35.08	—
26	30	50	5	All	0	0	0.41	0.42	—
27	50	30	5	All	180	205	11.38	13.23	—
28	50	50	2	All	472	576	51.42	60.79	—

TABLE 5. *Type 3 Data Sets*

Data set	m	$n+1$	Total problems	Total integer	Average pivots	Maximum pivots	Average time	Maximum time	Total LP subproblems
29	15	15	5	All	39	59	0.44	0.72	—
30	20	20	5	All	76	110	1.46	2.40	—
31	25	25	5	All	79	101	2.13	2.78	—
32	20	40	5	All	0	0	0.21	0.22	—
33	40	20	5	All	120	187	3.86	6.29	—
34	30	30	5	4	135	221	4.95	8.69	1
35	40	40	5	4	145	223	8.71	13.57	1
36	30	50	5	All	8	38	0.73	2.11	—
37	50	30	5	All	152	179	8.05	10.06	—
38	50	50	2	0	302	388	27.29	34.78	2

9. ACKNOWLEDGEMENT

We wish to thank D. S. Rubin for suggesting the column generation heuristic and for suggesting the procedure for storing the basis inverse, D. S. Dannenbring and B. M. Khumawala for the use of their data and computational results, and an anonymous referee for proposing a number of valuable revisions.

REFERENCES

[1] Dannenbring, D. G. and B. M. Khumawala, "An Investigation of Branch and Bound Methods for Solving the Segregated Storage Problem," AIIE Transactions, *5*, 265–274 (1973).

[2] Eilon, S. and R. V. Mallya, "A Note on the Optimal Division of a Container into Two Compartments," International J. Production Res. *5*, 1963–1969 (1966).

[3] Evans, J. R., "An Implicit Enumeration Algorithm for the Segregated Storage Problem," Working Paper, School of Industrial and Systems Engineering, Georgia Institute of Technology, (1974).

[4] Garfinkel, R. S. and G. L. Nemhauser, *Integer Programming* (John Wiley, 1972).

[5] Gilmore, P. C. and R. E. Gomory, "A Linear Programming Approach to the Cutting-Stock Problem," Operations Research *9*, 849–859 (1961).

[6] Hadley, G. F., *Nonlinear and Dynamic Programming* (Addison-Wesley, 1964).

[7] Lasdon, L. S., *Optimization Theory for Large Systems* (Macmillan, 1970).

[8] Magazine, M. J., G. L. Nemhauser, and L. E. Trotter, Jr., "When the Greedy Solution Solves a Class of Knapsack Problems," Operations Research *23*, 207–217 (1975).

[9] Shlifer, E. and P. Naor, "Elementary Theory of the Optimal Silo Storage Problem," Operational Research Quarterly *12*, 54 (1961).

[10] White, J. A., "Normative Models for Some Segregated Storage and Warehouse Sizing Problems," unpublished Ph. D. Dissertation, Ohio State University (1970).

[11] White, J. A. and R. L. Francis, "Solving a Segregated Storage Problem Using Branch-and-Bound and Extreme Point Ranking Methods," AIIE Transactions *3*, 37–44 (1971).

OPTIMAL FACILITY LOCATION UNDER RANDOM DEMAND WITH GENERAL COST STRUCTURE

V. Balachandran

and

Suresh Jain

Graduate School of Management
Northwestern University
Evanston, Illinois

ABSTRACT

This paper investigates the problem of determining the optimal location of plants, and their respective production and distribution levels, in order to meet demand at a finite number of centers. The possible locations of plants are restricted to a finite set of sites, and the demands are allowed to be random. The cost structure of operating a plant is dependent on its location and is assumed to be a piecewise linear function of the production level, though not necessarily concave or convex. The paper is organized in three parts. In the first part, a branch and bound procedure for the general piecewise linear cost problem is presented, assuming that the demand is known. In the second part, a solution procedure is presented for the case when the demand is random, assuming a linear cost of production. Finally, in the third part, a solution procedure is presented for the general problem utilizing the results of the earlier parts. Certain extensions, such as capacity expansion or reduction at existing plants, and geopolitical configuration constraints can be easily incorporated within this framework.

NTRODUCTION

The facility location problem and related problems are of relevance in the long range planning of m's operations. These problems involve the determination of the location of the facilities, their ciated capacity, and the distribution of the product from these facilities to the different demand ers. Different aspects of this problem have been investigated by a number of researchers under d assumptions [2, 10, 14, 15, 17]. In our paper we consider a generalized version of the above lem. Specifically, we consider the case where the location of facilities and their sizes are to be ded upon in order to satisfy the demand from different centers. The demand at these different ers is assumed to be random, and the cost associated with the production at any facility is assumed a piecewise linear function though not necessarily convex or concave. In the next section we con- the problem formulation and its motivation.

IODEL FORMULATION

A firm manufactures a product which is required at n different demand centers. The demand $= 1, 2, \ldots, n)$, at each center, is assumed to be a random variable whose marginal density $f(b_j)$ is med to be known. The firm has the option of setting up facilities at different sites, $i (i = 1, 2, \ldots, m)$. possible capacities of the facility at site i can be any nonnegative integer a_{ir} from the ordered $_i$, where $A_i = \{a_{ir} \mid r = 1,2, \ldots, m_i$ and $a_{ir} < a_{ir+1}$ for $r = 1,2, \ldots, m_i - 1\}$. The first element

of each of the sets A_i $(i = 1,2, \ldots, m)$ is assumed to be 0 and corresponds to the decision of not lo cating a facility at site i, and the last element a_{im_i} corresponds to the maximum possible productio at site i. The cost of producing y_i units at site i is $f_i(y_i)$ where $f_i(\cdot)$ is a piecewise linear function define on the set of nonnegative integers. Specifically,

$$(1) \qquad f_i(y_i) = \begin{cases} 0 & \text{if } y_i = a_{i1} = 0 \\ K_{ir} + v_{ir}y_i & \text{if } a_{ir} < y_i \leq a_{i,r+1} \text{ for } r = 1,2, \ldots, m_{i-1}. \\ \infty & \text{if } y_i > a_{im_i} \end{cases}$$

K_{ir} may be considered as the fixed component of the cost associated with setting up a plant of maximur capacity $a_{i,r+1}$, and v_{ir} is the per unit variable cost. The cost of distributing x_{ij} units from a facilit at site i to demand center j is $t_{ij}x_{ij}$ where t_{ij} is a constant. These costs may be considered as the dis counted costs if a multiperiod planning horizon is considered. With the above notation the problen can be formulated:

$$(2) \qquad \text{Minimize } Z = E\left[\sum_{i=1}^{m} f_i(y_i) + \sum_{i=1}^{m} \sum_{j=1}^{n} t_{ij}x_{ij} \right]$$

subject to

$$(3) \qquad \sum_{i=1}^{m} x_{ij} = b_{j*} \qquad \text{for } j \epsilon J,$$

$$(4) \qquad \sum_{j=1}^{n} x_{ij} = y_i \leq a_{im_i} \qquad \text{for } i \epsilon I,$$

$$(5) \qquad x_{ij} \geq 0,\ y_i \geq 0,$$

where $J = \{1, 2, \ldots, n\}$, $I = \{1, 2, \ldots, m\}$ and b_{j*} represents the realization of the random variable b_j

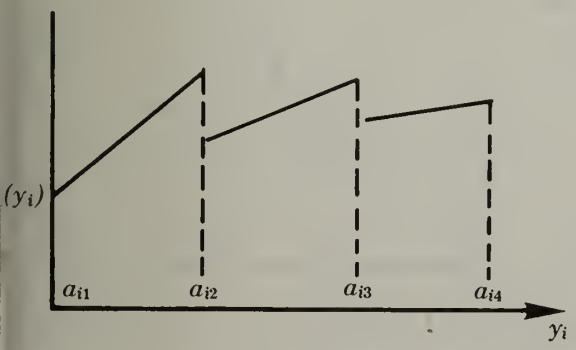
The above formulation incorporates the problems posed by various authors. For instance, the capacitated plant location problem [10], the location-allocation problem [14], and the warehouse loca tion problem [2, 17], are all special cases of the problem posed above. Discussions of the recent re search work done in this area have been presented by White and Francis [15], Soland [14] and recently by Elshafei and Haley [6]. Our formulation is similar in some respects to that of Soland [14]. Soland considers the case where the demand is deterministic and the cost functions are concave. Since the demand is generally not known with certainty as discussed in [7, 16], the consideration of this problem with probabilistic demand appears to be more realistic. White and Francis [15] consider probabilistic demand for a different problem, viz., that of determining optimum warehouse sizes only. The problem that they consider is different from ours, since they are not concerned with either the location, or the distribution aspects. Essentially their problem has no constraints of the type represented by Equations (3) and (4). The consideration of random demands increases the complexity of the problem. To our knowledge, no computationally satisfactory solution of this entire problem exists in the published literature.

Further, the cost structure that is generally considered is either a fixed cost plus a variable cost
ch that the total cost of production is concave, or a general concave cost as in Soland [14]. In the real
rld, because of indivisibilities and economies/diseconomies of scale, the cost structure is often
ferent. Hadley and Whitin [9, Chapter 2, p. 62] discuss quantity discounts where the cost function,
ugh piecewise linear, is neither concave nor convex. Rech and Barton [11] also consider nonconvex
cewise linear cost functions for solving the transportation and related network problems. Their
orithm, for deterministic problems, can be applied to the problem we treat in section 3 and is similar
the algorithm we develop there. Their algorithm is different, however, in that its underestimating
ictions are different, and it solves a sequence of network flow problems. By considering a sequence
transportation problems, we can use the operator theoretic approach to find the solutions without
olving. This use of transportation problems becomes even more important when the probabilistic
blem is considered later.

For ease of exposition we present the solution procedure of this problem in three phases. Initially
consider the case where the demand is deterministic. We develop an algorithm treating the b_{j*}'s
3) as known constants. This algorithm essentially solves a deterministic facility location, capacity,
l distribution problem with a general piecewise linear cost structure. A branch and bound procedure
erein subproblems are solved using operator theory [12, 13] (an extension of parametric programming
ere simultaneous changes in the parameters of a transportation problem are investigated) is pre-
ted in section 3 for the solution of this deterministic problem. In section 4 we develop an algorithm
olve the probabilistic case when the cost is assumed to be a linear function. The algorithm for this
babilistic demand problem utilizes the operator theoretic approach and the Kuhn-Tucker conditions.
ally, in section 5, we integrate the algorithms in sections 3 and 4 to develop a solution procedure
the entire problem. We show that the branch and bound procedure of section 3 for the entire problem
ults in subproblems of the type considered in section 4. This three-phase approach provides flexi-
y for a user to solve either the general problem or the special cases considered in sections 3 and 4.

THE DETERMINISTIC DEMAND MODEL

In this section, we assume that the b_{j*}'s given in Equation (3) are known constants and in order
nsure a feasible solution to this deterministic problem we assume that $\sum_{i=1}^{m} a_{im_i} \geq \sum_{j=1}^{n} b_{j*}$. Some of
different cost structures which arise in reality, and are permissible in our formulation, are sketched
igures 1a–1c.



URE 1a. Economies of scale both in fixed and variable cost.

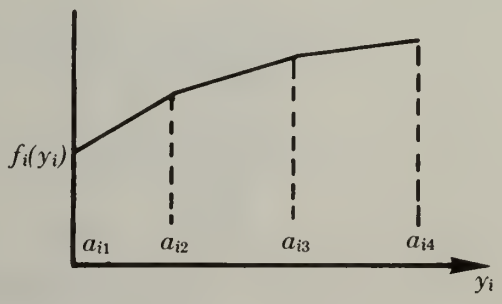


FIGURE 1b. Piecewise linear concave costs.

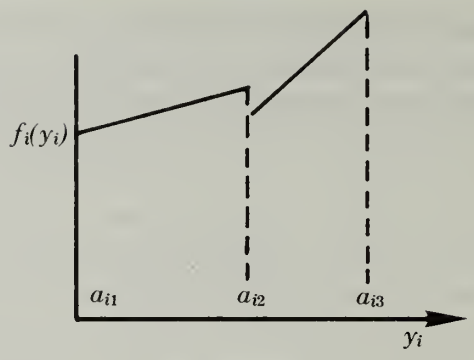


FIGURE 1c. Diseconomies of scale in variable cost.

To solve this problem we apply the branch and bound procedure. We first approximate all th cost functions, $f_i(y_i)$, by their best linear underestimates.

DEFINITION 1: A linear underestimate of the function $f_i(y_i)$ over the interval R_i is a line function $l_0 + l_i y_i$ such that

$$l_0 + l_i y_i \leq f_i(y_i) \text{ for all } y_i \epsilon R_i,$$

and

$$l_0 + l_i y_{i0} = f_i(y_{io}) \text{ where } y_{io} \epsilon R_i \text{ and } y_{i0} \leq y_i, \forall\, y_i \epsilon R_i.$$

DEFINITION 2: The best linear underestimate $c_0 + c_i y_i$ of the function $f_i(y_i)$ over the interval R_i is a linear underestimate such that if $l_0 + l_i y_i$ is any linear underestimate of the function $f_i(y_i)$ ov R_i then $c_i \geq l_i$.

As illustrated in Figure 2, OE is the best linear underestimate of the cost functions sketched in th interval $[0, a_3]$. The best linear underestimates of the cost function in the intervals $[0, a_2]$ and $[a_2, a$ are OB and CD respectively.

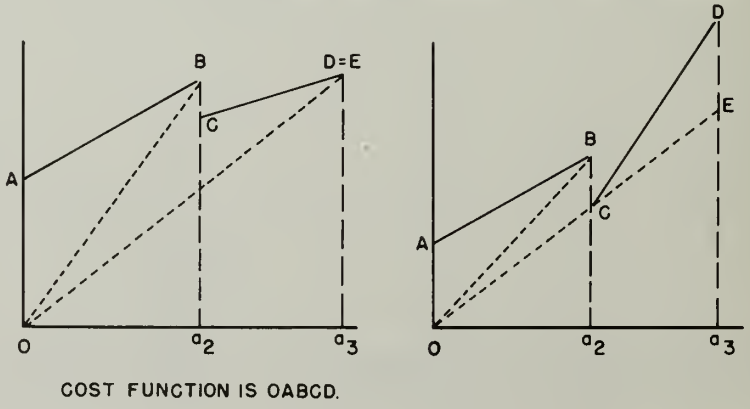


FIGURE 2. Illustration of the best linear underestimates.

Substituting the best initial linear underestimates $c_{i0}^1 + c_i^1 y_i$ for each of the function $f_i(y_i)$ over the rvals $[a_{i1}, a_{im_i}]$ we obtain the following transportation problem after removing the constant term, ny, from the objective function (corresponding to node 1 in Figure 3). Note that the constant term ero at node 1.

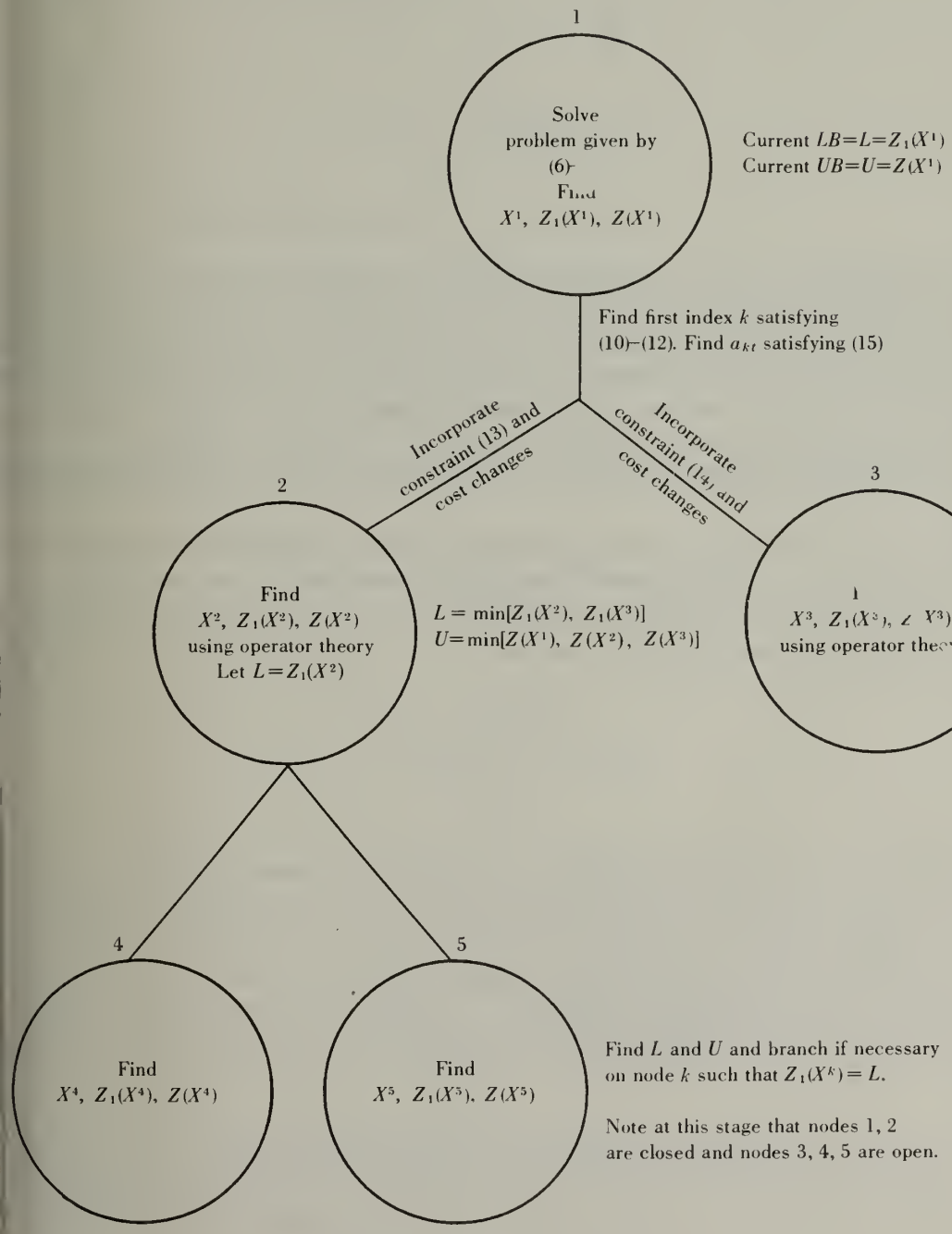


FIGURE 3. Branch and bound procedure.

(6) $$\text{Minimize } Z_1 = \sum_{i=1}^{m} \sum_{j=1}^{n} (c_i^1 + t_{ij}) x_{ij}$$

subject to

(7) $$\sum_{i=1}^{m} x_{ij} = b_{j*} \qquad \text{for } j \epsilon J,$$

(8) $$\sum_{j=1}^{n+1} x_{ij} = a_{im_i} \qquad \text{for } i \epsilon I,$$

(9) $$x_{ij} \geq 0 \qquad \text{for } i \epsilon I \text{ and } j \epsilon J \cup \{n+1\},$$

where, $x_{i,n+1}$ $(i = 1,2, \ldots,m)$ are the slack variables. Let us denote by J' the set $J \cup \{n+1\}$. l the optimal solution to this approximate problem be $X^1 = \{x_{ij}^1\}$ and its optimal cost be $Z_1(X^1)$. (general at node s we let X^s and y_i^s denote the optimal solution of the approximate transportati problem, and the resultant production at plant i respectively. Note that $y_i^s = \sum_{j \epsilon J} x_{ij}^s$.) If $Z(X^1)$ is t value of the original objective function (2) for the solution vector X^1 then we can easily prove the follo ing result:

If X^* is the optimal solution vector to the original problem represented by equations (2)–(5) then the optimal value $Z(X^*)$ is bounded above by $Z(X^1)$ and bounded below by $Z_1(X^1)$.

If L and U denote the current lower and upper bounds, then after solving the first approxima problem corresponding to node 1 (Figure 3), we have

$$L = Z_1(X^1),$$

and

$$U = Z(X^1).$$

After solving the first approximate problem, we partition the domain of definition of one of t functions $f_i(y_i)$ so as to obtain better linear underestimates in each of the different segments of t partition. A number of different rules could be used to determine the index $i(i \epsilon I)$ on which to par tion. Further, there is also the option of determining the number of segments in the partition, whi would equal the number of branches. In this paper we provide one rule for obtaining two branche This rule is similar to the one proposed by Rech and Barton [11]. We determine the first index k, k such that

(10) $$f_k(y_k^1) - c_k^1 y_k^1 - l_{k0}^1 \geq f_i(y_i^1) - c_i^1 y_i^1 - l_{i0}^1, \qquad \text{for } i \epsilon I,$$

and

(11) $$f_k(y_k^1) - c_k^1 y_k^1 - l_{k0}^1 > 0,$$

ere

$$y_i^1 = \sum_{j \in J} x_{ij}^1.$$

two branches that we obtain correspond to

$$y_k \leqslant a_{k,t},$$

$$y_k > a_{k,t},$$

ere

$$a_{k,t} < y_k^1 \leqslant a_{k,t+1}.$$

Thus at each stage of the branching process we generate two additional subproblems which can epresented by nodes in the branching tree. These nodes are numbered sequentially as the sub-blems are generated. These subproblems partition the domain of definition of $f_k(y_k)$. We substitute best linear approximation of $f_k(y_k)$ in each of these partitions and add the relevant constraint from set (13)–(14) to each of the subproblems. The resultant problems are capacitated transportation blems. In general, at a specific node s, which is obtained as a branch from node p by partitioning n index $k(k \epsilon I)$ satisfying inequalities similar to (10)–(11) the following problem results:

$$\text{Minimize } Z_1 = \sum_{i=1}^{m} \sum_{j=1}^{n} (c_i^s + t_{ij}) x_{ij} + K^s$$

ect to

$$\sum_{i=1}^{m} x_{ij} = b_{j*} \qquad \text{for } j \epsilon J,$$

$$\sum_{j=1}^{n+1} x_{ij} = U_i^s \qquad \text{for } i \epsilon I,$$

$$\begin{cases} x_{in+1} \leqslant u_i^s & \text{for } i \epsilon I, \\ x_{ij} \geqslant 0 & \text{for } i \epsilon I \text{ and } j \epsilon J', \end{cases}$$

e

(i) $c_i^s = c_i^p$, $U_i^s = U_i^p$, $u_i^s = u_i^p$ for $i \neq k$,

ii) and for $i = k$, if $\gamma_{0k}^s + \gamma_k^s y_k$ is the best linear underestimate in the relevant partitioned domain of y_k at node s, then

$$c_k^s = \gamma_k^s, \quad K^s = K^p + \gamma_{0k}^s - \gamma_{0k}^p.$$

Further, if node s corresponds to the branch associated with an inequality similar to (13), w have $U_k^s = a_{kt}$, $u_k^s = u_k^p$. If node s corresponds to the branch associated with an inequali similar to (14), then $U_k^s = U_k^p$, $u_k^s = U_k^p - (a_{kt}+1)$.

Note: Initially at node 1, we have $K^1=0$, $U_i^1 = a_{im_i}$, $u_i^1 = a_{im_i}$, and the c_i^1 have been defined earli

Thus the following changes occur in each of the branches:

(i) cost coefficients of all x_{kj}, $j=1, 2, \ldots, n$ differ by some constant amount $-k_1$ in one of t branches and k_2 in the other branch; and

(ii) capacity of plant k is changed in one of the branches, and the upper bound of the slack va able $(x_{k,n+1})$ corresponding to plant k is changed in the second one.

We may remark here that the new solutions, taking into consideration the above changes, can b generated (without resolving) by the operator theoretic approach discussed in [12, 13].

In Figure 3, if X^2 and X^3 are the optimal solutions to the approximate problems generated by th branching process at nodes 2 and 3, then we can show that the current lower bound, L, and upp bound, U, satisfy the following (by an argument similar to our earlier result):

$$L = \text{Minimum } [Z_1(X^2), Z_1(X^3)] \geq Z_1(X^1),$$

$$U = \text{Minimum } [Z(X^1), Z(X^2), Z(X^3)],$$

and

$$L \leq Z(X^*) \leq U.$$

REMARK: The strict inequality $L > Z_1(X^1)$ generally holds because the best linear underestimate in at least one of the partitions has to be strictly greater than the previous best linear underestimat since (11) holds. In the case when there is an alternate optimal solution and the revised best linea underestimate corresponding to the branch with this alternate optimal solution is unchanged, then th strict inequality will not hold.

If $L = Z_1(X^2)$ we branch on the node corresponding to the solution X^2 to obtain nodes 4 and 5 a shown in Figure 3. If X^4 and X^5 are the solutions at these nodes, then the new current lower and upp bounds are given by

$$L = \text{Min } [Z_1(X^4), Z_1(X^5), Z_1(X^3)],$$

and

$$U = \text{Min } [Z(X^1), Z(X^2), Z(X^3), Z(X^4), Z(X^4)].$$

Thus the current lower bound, L, at any stage equals the minimum of the lower bounds at the ope nodes (nodes from where there are no branches), whereas the current upper bound, U, is the minimu of the upper bounds over all the nodes. The algorithm terminates when the current lower bound equal the current upper bound at the same stage. This process terminates in a finite number of steps sinc the number of sites is finite and since, at each branch, we partition the domain of definition of the y_i into disjoint intervals. Further, since each of the $f_i(y_i)$ is a piecewise linear function, we cannot hav more than m_i partition on the variable y_i before we have the best linear underestimate equal the func tion itself, and thereby not allowing further branching on the variable y_i.

We therefore have the following algorithm.

ALGORITHM I:

(Step 0) For each of the functions $f_i(y_i)$ substitute $c_{i0}^l + c_{i}^l y_i$, the best linear underestimate of $f_i(y_i)$ $a_{i1} \leq y_i \leq a_{im_i}$. Solve the resultant transportation problem, and create an open node corresponding he solution X^1. The current lower bound is $Z_1(X^1)$ and the current upper bound is $Z(X^1)$.

(Step 1) If the current lower bound equals the current upper bound and occur at the same node n terminate. The optimal solution corresponds to the solution at the node where these bounds are al. Otherwise go to Step 2.

(Step 2) Determine the open node with the current lower bound and partition on the index k that sfies Equations (11)–(13). Close this node and generate two additional open nodes corresponding Equations (13) and (14). Solve the two resultant transportation problems using operator theory from solution of the old open node. If any of the resultant transportation problems has no feasible solu- then close that node and drop it from consideration for further branching. Let the upper bound ociated with such a node equal ∞. Determine the current lower bound over all the open nodes, and current upper bound over all the nodes. Go to Step 1.

REMARK: If at a branch (node) some $y_k \leq a_{k1} = 0$ (no plant at site k), then in the resultant trans- ation problem the variables x_{kj} for $j \in J'$ can be set to zero, or alternatively, dropped.

THE PROBABILISTIC MODEL

In this section we investigate the problem in Equations (6′)–(9′) with special emphasis on the case n the demands b_j, $j \in J$, are random. The demand is assumed to follow some known multivariate ribution, which allows interaction among the demands at the different centers. Let $f(b_j)$ represent marginal density function of b_j. We therefore have the following problem (after removing the super- ts for ease of exposition).

$$\text{Minimize} \sum_{i \in I} \sum_{j \in J} c_i x_{ij} + \sum_{i \in I} \sum_{j \in J} t_{ij} x_{ij} + K$$

that

$$\sum_{i \in I} x_{ij} = b_j \qquad \text{for } j \in J,$$

$$\sum_{j \in J} x_{ij} + x_{i,\, n+1} = U_i \qquad \text{for } i \in I,$$

$$x_{ij} \geq 0 \qquad \text{for } i \in I \text{ and } j \in J',$$

$$x_{i,\, n+1} \leq u_i \qquad \text{for } i \in I,$$

e each b_j is assumed to be random with known marginal density (or mass) function $f(b_j)$. Though a discrete random variable, for ease of theoretical exposition, we consider only the continuous ximation of b_j.

The solution procedure presented in this section is based on the theory of "Two Stage Linear amming Under Uncertainty" also known as "Stochastic Programming with Recourse", as pro-

posed by Dantzig and Madansky [5], and others, [4, 7, 8, 16]. In order to solve this problem we mak the following assumptions, which are similar to the ones used by all the above mentioned authors.

(A1) The marginal distribution $f(b_j)$ of each b_j is known.

(A2) This distribution $f(b_j)$ is independent of the choice of x_{ij}.

Charnes, Cooper and Thompson [4] have shown that the two-stage linear program is equivale to a constrained generalized median problem whose objective function has some absolute value term This objective function was shown to be equivalent to a mathematically tractable function by Garstk [8]. In this section we essentially follow the approach and notation given by Garstka [8].

Since each b_j is random, all the constraints given by Equation (17) need not be exactly met. W therefore assume that the firm experiences an opportunity cost of lost demand. This cost, $p_j \geq 0$, assumed to be linear and is treated as the per unit penalty cost of not satisfying demand at center Similarly, if we have more supply than the demand, then there may be a penalty due to holding, storin and obsolescence. Let us assume that this per unit penalty, $d_j \geq 0$, due to overproduction, is also linea at center j. Following Charnes, Cooper and Thompson [4], the total penalty costs associated with d mand exceeding production is:

$$\frac{1}{2} p_j \left\{ \left| b_j - \sum_{i \epsilon I} x_{ij} \right| + \left(b_j - \sum_{i \epsilon I} x_{ij} \right) \right\}. \tag{21}$$

It can be verified easily that if the production exceeds demand, viz., if $\sum_i x_{ij} > b_j$, then the abov penalty is zero. Similarly, the total penalty cost associated with excess production is

$$\frac{1}{2} d_j \left\{ \left| \sum_{i \epsilon I} x_{ij} - b_j \right| + \left(\sum_{i \epsilon I} x_{ij} - b_j \right) \right\}. \tag{22}$$

Following the approach of two-stage linear programming under uncertainty [4, 5, 7], we consider ou objective function to equal the expected value of the total production costs, distribution costs, and cost due to under and over production. The problem (16)–(20) is therefore equivalent to

$$\text{Minimize } E \left[\sum_{i \epsilon I} \sum_{j \epsilon J} c_i x_{ij} + \sum_{i \epsilon I} \sum_{j \epsilon J} t_{ij} x_{ij} \right. \tag{23}$$

$$+ \frac{1}{2} \sum_{j \epsilon J} d_j \left\{ \left| \sum_{i \epsilon I} x_{ij} - b_j \right| + \left(\sum_{i \epsilon I} x_{ij} - b_j \right) \right\}$$

$$\left. + \frac{1}{2} \sum_{j \epsilon J} p_j \left\{ \left| b_j - \sum_{i \epsilon I} x_{ij} \right| + \left(b_j - \sum x_{ij} \right) \right\} \right] + K$$

ject to

$$\sum_{j \in J} x_{ij} + x_{i,\,n+1} = U_i \qquad \text{for } i \in I,$$

$$x_{ij} \geq 0 \qquad \text{for } i \in I \text{ and } j \in J,$$

$$x_{i,\,n+1} \leq u_i \qquad \text{for } i \in I.$$

us denote $\sum_{i \in I} x_{ij}$ by b_{jo}, and $c_i + t_{ij}$ by C_{ij}, for $i \in I$ and $j \in J$. It has been shown by Garstka [8] that objective function given in (23) can be transformed into

$$\text{Min } F_1 + F_2 + K,$$

re

$$F_1 = \sum_{i \in I} \sum_{j \in J} \frac{1}{2} (2C_{ij} - p_j + d_j) x_{ij},$$

$$F_2 = \sum_{j \in J} (p_j + d_j) \int_{b_{j0}}^{b_{jm}} \left(b_j - \sum_{i \in I} x_{ij} \right) f(b_j)\, db_j$$

re b_{jm} is the median of the random variable b_j whose marginal density (mass) function is $f(b_j)$. It is seen that F_1 is linear and it can be proved easily that F_2 is convex (see Garstka [8, page 11]). and μ_i are the dual variables corresponding to constraints (24) and (26) respectively, then by the n-Tucker conditions the optimal solution $X^* = \{x_{ij}^*\}$, μ_i^*, and λ_i^* satisfy the following:

$$\lambda_i^* \text{ are unconstrained for } i \in I$$

$$\mu_{i.}^* \leq 0 \text{ for } i \in I$$

$$C_{ij} - \frac{1}{2} p_j + \frac{1}{2} d_j - \sum_{j \in J} (p_j + d_j) \int_{b_{j0}}^{b_{jm}} f(b_j)\, db_j - \lambda_i^* \geq 0 \qquad \text{for } i \in I \text{ and } j \in J,$$

$$\sum_{j \in J'} x_{ij}^* = U \qquad \text{for } i \in I,$$

$$x_{i,\,n+1}^* \leq u_i \qquad \text{for } i \in I,$$

$$x_{ij}^* \geq 0 \qquad \text{for } i \in I \text{ and } j \in J',$$

$$x_{ij}^* \cdot \{\text{left hand side of (32)}\} = 0 \qquad \text{for } i \in I \text{ and } j \in J.$$

$$x_{i,\,n+1}^* (-\mu_i^* - \lambda_i^*) = 0 \qquad \text{for } i \in I,$$

$$(u_i - x^*_{i,n+1})\mu^*_i = 0 \qquad \text{for } i \in I. \tag{38}$$

These conditions given by (30)–(38) provide a basis for solving the stochastic capacitated transportation problem. Based on these conditions, the following propositions P1, P2, and P3 can be proved. The proofs are similar to those provided by Garstka [8].

P1. For any $i \in I$ and $j \in J$, if $C_{ij} > p_j$, then $x_{ij} \equiv 0$.

P2. For any $j \in J$ and all $i \in I$, if $C_{ij} \geq (p_j - d_j)/2$, then $\sum_{i \in I} x_{ij} \leq b_{jm}$ in the optimal solution. It may be noticed that if $p_j < d_j$, the above inequality trivially holds since $C_{ij} \geq 0$.

P3. For any $i \in I$ and all $j \in J$, if $C_{ij} + d_j < 0$ then $\sum_{j \in j} x_{ij} = U_i$.

In many situations the combined cost of production and transportation, C_{ij}, is greater than half the difference $(p_j - d_j)$. Thus in the algorithm given below we assume that $C_{ij} \geq (p_j - d_j)/2$. Note however, that if the per unit cost of underproduction, p_j, is less than that of overproduction, d_j, then the above assumption is unnecessary. In order to solve (30)–(38) let us pose the new deterministic transportation problem given below:

$$\text{Minimize} \qquad F_1 = \sum_{i \in I} \sum_{j \in J} \frac{1}{2} (2C_{ij} - p_j + d_j) x_{ij} \tag{39}$$

such that

$$\sum_{j \in J'} x_{ij} = U_i \qquad \text{for } i \in I \tag{40}$$

$$\sum_{i \in I} x_{ij} = b_{j*} \qquad \text{for } j \in J \tag{41}$$

$$x_{ij} \geq 0 \qquad \text{for } i \in I \text{ and } j \in J \text{ and} \tag{42}$$

$$x_{i,n+1} \leq u_i \qquad \text{for } i \in I \tag{43}$$

where b_{j*} are the realizations of b_j. If λ_i, μ_i, and ν_j are the dual variables corresponding to constraint sets (40), (43), and (41) respectively, then by the Kuhn-Tucker conditions the optimal primal and dual variables satisfy:

$$\lambda^*_i, \nu^*_j \text{ are unconstrained, and } \mu^*_i \leq 0, \qquad \text{for } i \in I, j \in J, \tag{44}$$

$$C_{ij} - (p_j + d_j)/2 - \lambda^*_i - \nu^*_j \geq 0, \qquad \text{for } i \in I, j \in J, \tag{45}$$

$$\sum_{j \in J'} x^*_{ij} = U_i, \qquad \text{for } i \in I, \tag{46}$$

$$\sum_{i \in I} x^*_{ij} = b_{j*}, \qquad \text{for } j \in J, \tag{47}$$

$$x^*_{i,n+1} \leq u_i, \qquad \text{for } i \in I,$$

$$x^*_{ij} \geq 0, \qquad \text{for } i \in I \text{ and } j \in J',$$

$$x^*_{ij} \cdot (\text{left hand side of (45)}) = 0 \qquad \text{for } i \in I \text{ and } j \in J,$$

$$x^*_{i,n+1}(-\mu^*_i - \lambda^*_i) = 0 \qquad \text{for } i \in I,$$

$$(u_i - x^*_{i,n+1})\mu^*_i = 0 \qquad \text{for } i \in I.$$

It is easy to observe the similarity of the Kuhn-Tucker conditions given in (44)–(52) to those of original problem. Comparing Equations (32) and (45), we see that a solution to (44)–(52) will satisfy conditions of the original problem (30)–(38) if

$$\nu^*_j = (p_j + d_j) \int_{b_{j0}}^{b_{jm}} f(b_j)db_j.$$

Since, by assumption $C_{ij} \geq (p_j - d_j)/2$, by proposition P2 stated earlier

$$b_{j0} = \sum_{i \in I} x_{ij} \leq b_{jm},$$

therefore the problem reduces to finding a set of b_j , $j \in J$, such that

$$\nu^*_j = (p_j + d_j) \int_{b_{j*}}^{b_{jm}} f(b_j)db_j$$

If such b_{j*} exist, then an optimal solution to (16)–(20) will be obtained by solving (39)–(43) with b_{j*} replacing b_j in the constraint set (17). An optimal solution can now be determined by the following ithm, under the assumption that $C_{ij} \geq (p_j - d_j)/2$, for every i and j.

ALGORITHM II:

(Step 0) *Initialization:* Let k the number of iterations be 1. Find b_{jm} the median of the random ble b_j for $\forall\ j \in J$, and set the initial $b_{j*} = b^1_{jm}$ for $j \in J$ in Equation (41). Find the optimal solution and nal cost to the deterministic transportation problem (39)–(43) and find the dual variables λ_i, μ_i $\in I$ and ν_j for $j \in J$. (Though μ_i for $i \in I$ are obtained they are not directly needed.) The duals λ_i and ν_j olved from the relation $\lambda_i + \nu_j = C'_{ij} = C_{ij} - (p_j - d_j)/2$ for the (i,j) in the optimal basis. For $k=1$ e basis set be denoted by $B^1 = \{(i,j) \mid x_{ij} > 0\}$ and $\Lambda^1 = \{\lambda^1_i \mid$ the optimal values of the dual varia- for $i \in I\}$ and $V^1 = \{\nu^1_j \mid$ the optimal value of the dual variables for $j \in J\}$.

(Step 1) *Iteration Procedure:* Find b^{k+1}_{j*} from the following (newsboy type) relationship:

$$\nu^k_j = (p_j + d_j) \int_{b^{k+1}_{j*}}^{b_{jm}} f(b_j)db_j.$$

(Step 2) If $b_{j^*}^{k+1} = b_{j^*}^{k}$ for *every* $j \epsilon J$, then an optimal solution to the stochastic transportation proble (16)–(20) is found and hence STOP, otherwise go to Step 3.

(Step 3) *Area Rim Operator Application:* Let $\beta_j^{k+1} = b_{j^*}^{k+1}$, for $j \epsilon J$ and let $\alpha_i = 0$ for $i \epsilon I$. Followi the algorithm given by Srinivasan and Thompson [12, page 215], apply the area operator δR^A with $\delta =$ and the above computed α_i and β_j to generate the new optimal solution for the $(k+1)$th problem. L the new dual vector be Λ^{k+1}. Let $k = k+1$ and go to Step (1). (Note that if V^{k+1} and V^k are the san then $b_{j^*}^{k+1} = b_{j^*}^{k}$ so that the iteration can be stopped.)

The convergence proof of this algorithm is based on the results derived by Charnes and Coop [3] and Charnes, Cooper, Thompson [4]. Garstka [8] has given a proof based on [3, 4] and since the pro for our algorithm is similar to that given by Garstka [8], we only provide an outline of the proof. It c be shown that $F_2(b_j)$ is convex in b_j so that $F_1 + F_2$ in (27) is also convex. Further, it can be shown th $b_{j^*}^{k+1} \leq b_{j^*}^{k}$, and thus the $b_{j^*}^{k}$ are monotone and bounded by $b_{j^*}^{1} = b_{jm}$. With the relationship betwe $b_{j^*}^{k+1}$ and ν_j^k which is one to one, and due to monotonicity and convexity, convergence is established.

5. THE GENERAL MODEL AND EXTENSIONS

In this section we provide a solution procedure for the problem formulated in section 2 given Equations (2)–(5). In section 3 we provided Algorithm I based on a branch and bound procedure, solve the above problem with the b_j's being deterministic. At every branch in this algorithm, we a faced with a deterministic transportation problem, where the right hand sides in the constraint s (3) and (4) change from branch to branch and the cost coefficients in (6) also change. However, due the operator theory [12, 13], the optimal solutions at each branch are obtained in a computational efficient manner. Let us represent, for ease of exposition, the costs and the right hand sides of a bran s as c_i^s, U_i^s, and u_i^s for $i \epsilon I$. Now, to consider the problem (2)–(5) in its entirety, it is enough if we intr duce the randomness in the b_j's, $j \epsilon J$. This leads us directly into section 4, where these appropria costs and right hand sides replace the corresponding ones in Equations (16)–(20). Thus, Algorithm I is directly applicable to this new stochastic transportation problem provided Assumptions A1–A hold. It is to be noticed that the t_{ij}, p_j and d_j do not change. Assumptions A1 and A2 are unrestricti and can be expected to be true in most practical situations. We now present the following unified A gorithm III to solve the original problem posed by Equations (2)–(6) by utilizing Algorithm I first ar applying Algorithm II to each branch of Algorithm I.

ALGORITHM III:

(Step 0) For each of the functions $f_i(y_i)$ substitute $c_{i0}^1 + c_i^1 y_i$, the best linear underestimate $f_i(y_i)$ for $a_{i1} \leq y_i \leq a_{im_i}$. Solve the resultant stochastic transportation problem (6)–(9) with the assum tion that b_j^* for $j \epsilon J$ in (7) is random, utilizing Algorithm II. Create an open node corresponding to th optimal solution X^1 to the stochastic transportation problem. Denote the current lower bound $Z_1(X^1)$ and from the objective function Z given in (23), get the upper bound $Z(X^1)$.

(Step 1) Same as Step 1 of Algorithm I.

(Step 2) Determine that open node with the current lower bound and partition on the index that satisfies Equations (10)–(13). Close this open node and generate two new open nodes as branche corresponding to (13) and (14). Note that the new problem at each of these branches is a stochast transportation problem which can be solved using Algorithm II. Determine the current lower bou over all the open nodes, and the current upper bound over all the nodes. Go to Step 1.

REMARK: Since Algorithm III above is the unification of Algorithms I and II, the convergence ows due to the convergence of the earlier two algorithms.

Our algorithms facilitate easy consideration of certain extensions to the problem formulated in ion 2 such as:

(1) Inclusion of constraints requiring mandatory operation of certain plants.

(2) Capacity expansion or reduction at existing plants.

(3) Geopolitical consideration requiring the operation of plants at mutually exclusive or mutually dependent plant sites.

Extensions (1) and (2) follow from the fact that each a_{im} is arbitrary and each a_{i1} can be made itrary, while extension (3) can be imposed when branching occurs.

ACKNOWLEDGEMENT

The authors wish to express their appreciation for the helpful comments of a referee.

BIBLIOGRAPHY

Balachandran, V., "The Stochastic Generalized Transportation Problem—An Operator Theoretic Approach," Management Science Research Report No. 311, GSIA, Carnegie-Mellon University, Pittsburgh, Pa., (Feb. 1973).

Baumol, W. J. and P. Wolfe, "A Warehouse Location Problem," Operations Research *6*, 252–263 (1958).

Charnes, A. and W. W. Cooper, "Systems Evaluation and Repricing Theorems," Management Sci. *9*, 33–49 (1962).

Charnes, A., W. W. Cooper, and G. L. Thompson, "Constrained Generalized Medians and Hypermedians as Deterministic Equivalents for Two-Stage Linear Programs Under Uncertainty," Man. Sci. *12*, 83–112 (1965).

Dantzig, G. B. and Albert Madansky, "On the Solution of Two-Stage Linear Programs Under Uncertainty," Proceedings Fourth Berkeley Symposium on Mathematical Statistics and Probability, *1*, J. Neyman, ed., Berkeley (1961).

El Shafei, A. N. and K. B. Haley, "Facilities Location: Some Foundations, Methods of Solution, Application and Computational Experience," Operations Research Report No. 91, North Carolina State University, Raleigh (1974).

Ferguson, A. R. and G. B. Dantzig, "The Allocation of Aircraft to Routes—An Example of Linear Programming Under Uncertain Demand," Management Science *3*, 45–73 (1956).

Garstka, S. J., "Computation in Stochastic Programs with Recourse," (Ph. D. Dissertation, Management Science Research Report No. 204, Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, Pa. (1970).

Hadley, G. and T. M. Whitin, *Analysis of Inventory Systems* (Prentice-Hall, Englewood Cliffs, N.J., 1963).

Manne, Alan S., "Plant Location Under Economies of Scale—Decentralization and Computation," Management Science *11*, 213–235 (1964).

Mech, P. and L. G. Barton, "A Non-Convex Transportation Algorithm" in *Applications of Mathematical Programming Techniques*, E. M. L. Beale, (ed.), American Elsevier Publishing Co., New York, 1970), pp. 250–260.

[12] Srinivasan, V. and G. L. Thompson, "An Operator Theory of Parametric Programming for t Transportation Problem—I," Nav. Res. Log. Quart. *19*, 205–226 (1972).

[13] Srinivasan, V. and G. L. Thompson, "An Operator Theory of Parametric Programming for t Transportation Problem—II," Nav. Res. Log. Quart. *19*, 227–252 (1972).

[14] Soland, Richard M., "Optimal Facility Location with Concave Costs," Operations Research, 2 373–383 (1974).

[15] White, John A. and R. L. Francis, "Normative Models for Some Warehouse Sizing Problems AIIE Transactions *3*, No. 3, 185–190 (1971).

[16] Williams, A. C., "A Stochastic Transportation Problem," Operations Research *11*, 759–770 (1963)

[17] Zoutendijk, G., "Mixed Integer Programming and the Warehouse Allocation Problem," in *Applic tions of Mathematical Programming Techniques*, E. M. L. Beale, (ed.) (American Elsevi Publishing Co., New York, 1970) pp. 203–215.

SOME COMPARATIVE & DESIGN ASPECTS OF FIXED CYCLE PRODUCTION SYSTEMS

A. L. Soyster

Temple University
Philadelphia, Pa.

D. I. Toof

Arthur Young & Co.
Washington, D.C.

ABSTRACT

A serial production line is defined wherein a unit is produced if, and only if, all machines are functioning. A single buffer stock with finite capacity is to be placed immediately after one of the first $N-1$ machines in the N machine line. When all machines have equal probability of failure it is shown that the optimal buffer position is exactly in the middle of the line. This result is synthesized with the earlier work of Koenigsberg and Buzacott including an analysis of the covariance between transition states. An alternative model formulation is presented and integrated with previous results. Finally, a sufficient condition and solution procedure is derived for the installation of a buffer where there is a possible trade-off between increasing the reliability of the line versus adding a buffer stock.

NTRODUCTION

One of the earlier problems considered in the operations research literature is the analysis of in types of production lines. In an often cited reference [13] Koenigsberg states that three major lems in the design and operation of production lines are (a) the number of stages in the line, (b) the ion of the buffer stocks, and (c) the capacity of these buffer stocks. Buzacott [4, 5, 6, 7] has stigated these problems under various assumptions of machine reliability and product flow. For a t survey of the pertinent literature see Buxey et al. [3]. In this paper we shall be concerned only the latter two questions in the context of a fixed cycle serial production line. In section 5 we com- and integrate our results with those of Koenigsberg and Buzacott.

The term production line or automatic production line is a concept that is, in general terms, familiar nost everyone. However, much of the previous literature has not adequately described the actual anism of product flow within the production facility. One considers a sequence of facilities spe- lly designed to convert raw materials into certain finished goods. In [13] Koenigsberg specifies in general descriptors of production lines. A typical model of a production line could be as in te 1.

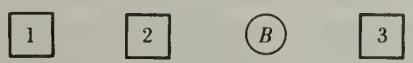


FIGURE 1

Here one has a production line with three stages and an allocation of space for a buffer stock betw stages 2 and 3. The output of stage 1 is the input of stage 2, while the output of stage 2 is the inpu stage 3. The buffer stock is used as follows: if some part of the production line prior to the buffer fa then the semifinished buffer units are available for the operation of the latter part of the production l The buffer acts to decouple the production line.

In view of Figure 1 the three general questions posed by Koenigsberg seem reasonable, bu number of fundamental questions concerning the actual mechanics of the product flow must be sp ified. Answers to these questions will depend upon factors such as cycle time, transfer times, down t associated with each stage and assumptions concerning a deterministic versus stochastic versio these parameters. One must specify the manner in which the product is transferred from the star the line to the last stage of the production line. One must also specify how units arrive for proces at the first stage of the production line.

One convenient dichotomy of production lines is whether units arrive at stage 1 in a ranc manner or in some deterministic fashion. Actually this dichotomy is more specific, as it is gener assumed that either units arrive in a Poisson fashion at stage 1 or that units are always available w needed at stage 1, i.e., an infinite supply. One of the earliest analytical models of automatic product lines is for the following general case in Figure 2.

FIGURE 2

Suppose that the capacity of each buffer B_i is infinite and that units arrive at stage 1 in a Pois fashion and, furthermore, that the actual time to complete processing of a unit at each stage is an ponential random variable. Burke [2] then shows that the departure distribution of stage 1 is actu Poisson. Hence the input rate to stage 2 is again Poisson, and the same argument can be repea concerning the output distribution of stage 2, etc. These assumptions concerning Poisson arrivals a exponential processing times at each stage lead in general to very tractable analytical results c cerning the system's parameters. For example, in [12] Hunt generates the mean number of units the system. When there is a finite capacity on the buffer size, the analysis of certain characteristic this sequential array of queues becomes considerably more complex, see e.g., [11].

In this paper we shall not take the approach that units arrive randomly at the first stage. The m realistic assumption is that units are available at the first stage whenever they are needed. Such assumption in general is the appropriate one for the typical production line. In addition, the assumpt is made that the processing time at each stage is a constant, i.e., fixed cycle production lines. A cert element of uncertainty, though, is introduced via the reliability of each stage.

Along with other defining characteristics, Koenigsberg makes this dichotomy of random ver deterministic arrival of units at stage 1. The two approaches are coined queuing and stochastic in [1 This paper would be correctly labeled the stochastic case. In the next section a definition of a se production line is proposed. Along with the definition of a serial production line an alternative prod tion line model is defined for comparison purposes. This alternative model underscores the need accurately define the micromechanics of how units are to be handled and transferred along the prod tion line. Apparently such considerations seem to attract little attention in the literature.

ESCRIPTION OF AN EXACT SERIAL MODEL

Consider an N stage production line where the output of machine i is the input to machine $i+1$. Let T_i, a constant be the processing time of machine i and define the cycle time, $T=\sum_{i=1}^{N} T_i$. During each cycle each machine i has a known probability p_i, of failing. The probability of all hines working in a given cycle is $\prod_{i=1}^{N} q_i$, where $q_i=1-p_i$.

Raw material units from an unlimited supply are available as input to machine 1 and finished s are the output of machine N. We shall assume that the machines in the production line behave ogously to a set of lights connected in a *series* circuit, i.e., if one or more of the machines fail the production line fails. At the beginning of each cycle a check is made to determine the status ch machine. If each of the machines is up (not failed), a raw material unit is processed through the and the probability that all machines will be up is $\prod_{i=1}^{N} q_i$. The distribution of up cycles of machine i ometrically distributed with parameter q_i, and the probability of successfully completing a unit in given cycle is $\prod_{i=1}^{N} q_i$. Note that during each cycle the failure probability, p_i, of machine i, is constant, we assume a constant repair time of one cycle. This, of course, is equivalent to the assumption he distribution of repair time, measured in cycles, is geometric with parameter p_i. In a subsequent on we consider the serial model when the distribution of repair time is of a more general form. In cular we shall explore some of the implications when the status of machine i, i.e., up or down, is ndent upon its status during the previous cycle.

Now suppose that there exists an operator associated with each machine i. Consider what this model implies as to the mechanics of how the semifinished units are transferred throughout the Suppose that at the beginning of a particular cycle all machines are up. In this case operator 1 sses a raw material unit at machine 1 and then transfers the semifinished unit to operator 2. operator 2 processes the semifinished unit apparently all the operators 1, 2, 4, . . . N remain Note that during a cycle only one semifinished unit is ever handled on the production line. If at eginning of a cycle one or more machines are down (failed), then this cycle is aborted and the line ns idle during the cycle time, T.

n section 5, we demonstrate that this definition is implicit in much of the earlier literature concern- oduction line models.

he applicability of this serial model would depend upon a number of factors in any real world ion. One application should include the case wherein a single worker (computer, another machine) tes the entire line, i.e., moves from machine to machine with the semifinished unit.

ext consider a simple production line with two machines, but between the two machines a buffer capacity of M units is installed (Figure 3).

FIGURE 3

The buffer consists of units that have already been processed on machine 1. Consider a cycle where machine 1 is down, but machine 2 is up. If the buffer is not empty, then a unit from the buffer can processed through machine 2. The cycle is successful in the sense that a finished good unit is obtain from the production line.

Next consider a cycle wherein machine 2 is down, but machine 1 is up. If the number of units the buffer is less than M a raw material unit can be processed through machine 1 and placed in t buffer. If the buffer is already at capacity no such transfer can take place. In either case no finished go units for this cycle are discharged from the production line. Note that the size of the buffer stock v vary between zero and M.

For the N machine line the question arises as to the optimal placement of a single buffer capaci for there would be $N-1$ possible positions. This question has earlier been considered by Buzacott [in the context of certain approximating models of fixed cycle production lines. In the following secti we develop a mathematical model of this serial production line and focus on the determination of t optimal placement of a single buffer capacity along the N machine line. The optimal placement will that position which maximizes the steady state probability that a given cycle furnishes a finished go unit.

Before proceeding to the next section, we contrast the assumptions of the serial model with anoth model. In [12] Hunt defines the "unpaced belt production line" as a particular case of a network queues. Actually, in this case he permits no queues at all, except for an infinite capacity queue prior the first machine, and furthermore, no vacant machines are allowed. Here the line is to move all at on This definition is in the context of a network of queues with Poisson arrivals and exponential servi times. For the case in which units are always available at machine 1 when needed, unit processing tim are constant, and when finite capacity buffers are permitted, a similar model can be defined. By t "sequential relay" model with finite buffer capacities, we shall interpret the schematic of Figure 3 the following manner: Each machine has a separate operator, and during each cycle (here defined the maximum of the individual unit processing times, T_i, That is, $T=\max_i T_i$), each operator adhe to the following rules:

(1) Operator 2 first:

(i) checks whether machine 2 is up;

(ii) if machine 2 is up, he checks whether an additional unit can be placed in the output buf (which we assume has infinite capacity for the last machine in the line);

(iii) if the output buffer can accept another unit, he checks to see if there is a unit availal from the input buffer.

If (i), (ii) and (iii) are all satisfied then operator 2 processes a unit on machine 2. Otherwise operato remains idle during this cycle. We assume the repair time of machine 2 is T, the cycle time.

(2) Operator 1 then:

(i) checks whether machine 1 is up;

(ii) if machine 1 is up, checks whether an additional unit can be placed in the output buff

(iii) if the output buffer can accept another unit, he checks to see if there is a unit availal from the input buffer (which we assume has an infinite supply for the first machine in t line).

If (i), (ii) and (iii) are all satisfied, then operator 1 processes a unit on machine 1. Otherwise operato remains idle during this cycle. Similarly, we assume the repair time for the machine 1 is T, the cy time.

For an N-machine line these rules can be expanded. The name sequential relay is used to accentu- the principle that, in this model, before operator i can determine his status during a cycle, operator 1 must first discern his status. Note that in the sequential relay model during any given cycle as ny as N distinct units can be processed on the N machines. An important contrast between the two dels is exhibited by Figure 3 when both machines are up and the buffer is empty. In the serial del a finished good unit is produced, while in the sequential relay model only machine 1 would be rated, and then at the start of the next cycle the buffer would contain a single unit.

A similar question concerning the optimal placement of a single buffer capacity is applicable to sequential relay model, although the analysis presented in the next section for the serial model is parently more complex for the sequential relay model. The difficulty lies in the fact that, by the very inition of the sequential relay model, a buffer of size one must be associated with each of the N chines.

MATHEMATICAL FORMULATION OF A SERIAL PRODUCTION LINE

For a serial production line with N machines there are $N-1$ possible positions for the buffer. The fer can be placed after the first machine, after the second, after the third, ..., and, lastly, after $N-1$ machine. Consider now the general case where the buffer with capacity M is placed after ith machine, $1 \leq i \leq N-1$. We shall demonstrate that the number of units in the buffer define a e Markov chain. Let $\alpha_i = \prod_{j=1}^{i} q_j$, $\beta_i = \prod_{j=i+1}^{N} q_j$, and let X_n be the number of units in the buffer at the inning of cycle n. Now consider the probability distributions for X_{n+1} for the following three cases:

(a) $X_n = 0$;

(b) $0 < X_n < M$;

(c) $X_n = M$.

case (a) $X_{n+1} = 0$ if

(i) All machines up during cycle n

(ii) At least one machine down before the buffer.

ce

$$P[X_{n+1} = 0 \mid X_n = 0] = \alpha_i\beta_i + (1-\alpha_i),$$

$$P[X_{n+1} = 1 \mid X_n = 0] = 1 - [\alpha_i\beta_i + (1-\alpha_i)] = \alpha_i - \alpha_i\beta_i.$$

Similar analysis applies to cases (b) and (c), and the complete transition matrix for this finite, ucible Markov chain given in Figure 4.

We now let $X^{(i)}$ denote the steady state distribution of the buffer size when the buffer is placed ediately after machine i. Define $\lambda_i = \alpha_i - \alpha_i\beta_i$, $\mu_i = \beta_i - \alpha_i\beta_i$ and $\rho_i = \lambda_i/\mu_i$. With this substitution ransition matrix portrays a well-known form whose steady state solution (see [8]) is

$$(1) \qquad P[X^{(i)}=t]=\begin{cases}\left(\dfrac{1-\rho_i}{1-\rho_i^{M+1}}\right)\rho_i^t & \lambda_i \neq \mu_i \\ \dfrac{1}{M+1} & \lambda_i=\mu_i\end{cases}$$

for $0 \leq t \leq M$. With the steady state solution for $X^{(i)}$ known, it is now possible to determine the stead state probability that a given cycle furnishes a finished good unit from machine N:

P[finished good when buffer placed between machines i and $i+1$]

$=P$[All machines up after the buffer] $\cdot$ {P[All machines up before the buffer]

$+P$[At least one machine down before the buffer] $\cdot P[X^{(i)} \geq 1]$}.

This quantity would be

$$\beta_i\{\alpha_i+(1-\alpha_i)P[X^{(i)} \geq 1]\}.$$

Note that

$$(2) \qquad P[X^{(i)} \geq 1]=\begin{cases}\dfrac{\rho_i-\rho_i^{M+1}}{1-\rho_i^{M+1}} & \lambda_i \neq \mu_i \\ \dfrac{M}{M+1} & \lambda_i=\mu_i\end{cases}$$

	0	1	2	. . .	$t-1$	t	$t+1$	. . .	$M-1$	M
0	$1-\alpha_i+\alpha_i\beta_i$	$\alpha_i-\alpha_i\beta_i$								
1	$\beta_i-\alpha_i\beta_i$	$1-\alpha_i-\beta_i+2\alpha_i\beta_i$	$\alpha_i-\alpha_i\beta_i$							
2		$\beta_i-\alpha_i\beta_i$	$1-\alpha_i-\beta_i+2\alpha_i\beta_i$							
.										
.										
.										
t					$\beta_i-\alpha_i\beta_i$	$1-\alpha_i\beta_i+2\alpha_i\beta_i$	$\alpha_i-\alpha_i\beta_i$			
.										
.										
.										
M									$\beta_i-\alpha_i\beta_i$	$1-\beta_i+\alpha_i\beta_i$

FIGURE 4. Transition matrix for buffer states.

nce to find the optimal placement of this single buffer with capacity M, one needs that particular $\leq i \leq N-1$, which maximizes

$$\beta_i\{\alpha_i + (1-\alpha_i)P[X^{(i)} \geq 1]\}.$$

Now consider the simpler case of (3) when $q_i = q$ for all i. In this special case it follows that $\alpha_i = q^i$, $= q^{N-i}$ so that $\lambda_i = q^i - q^N$, $\mu_i = q^{N-i} - q^N$ and

$$\rho_i = \frac{q^i - q^N}{q^{N-i} - q^N}.$$

te that if the buffer is placed immediately after machine $N/2$, i.e., $i = N/2$, then $\lambda_i = \mu_i$. Hence

$$P[X^{(N/2)} \geq 1] = \frac{M}{M+1}.$$

$\neq N/2$, then $\lambda_i \neq u_i$ and

$$P[X^{(i)} \geq 1] = \frac{\rho_i - \rho_i^{M+1}}{1-\rho_i^{M+1}}.$$

We now compute the probability for a given cycle of obtaining a unit from machine N when the fer is placed immediately after machine $N/2$. This probability from (3) is

$$q^{N-N/2}\left[\, q^{N/2} + (1-q^{N/2})\frac{M}{M+1}\,\right]$$

$$q^N + (q^{N/2} - q^N)\frac{M}{M+1}.$$

e that as M becomes large this probability approaches $q^{N/2}$, i.e., the second part of the line becomes pendent of the first $N/2$ machines.

Next consider the general form for the probability of obtaining a unit from machine N when the er is placed after machine $i \neq N/2$. Again from (3) and (5) one obtains

$$q^{N-i}\left[q^i + (1-q^i)\,\frac{\rho_i - \rho_i^{M+1}}{1-\rho_i^{M+1}}\right],$$

after simplifying one obtains as the steady state probability, ssp(i),

$$\text{ssp(i)} = q^N + \mu_i \frac{\rho_i^{M+1} - \rho_i}{\rho_i^{M+1} - 1}, \qquad i \neq N/2.$$

We shall show that for all integer $i \neq N/2$, $1 \leq i \leq N-1$, that (7) $<$ (6); in fact we shall show that if $i \in [1, N/2) \cup (N/2, N]$, then (7) $<$ (6). Hence from hereon we shall treat i as a continuous variable. The general strategy will be to show that the derivative of (7) with respect to i is greater than zero for $i \in [1, N/2)$ and less than zero for $i \in (N/2, N]$.

The derivative of (7) is obtained using the chain rule, and after simplification yields

(8)

$$\frac{lnq}{(\rho^{M+1}-1)^2}\{[\rho^M(\rho-1)(M+1)-(\rho^{M+1}-1)]q^i-\rho^M[(\rho^{M+1}-1)-(\rho-1)(M+1)]\rho q^{N-1}\}.$$

Bernoulli's inequality [1] states that if $\rho > 0$, then

$$(\rho^{M+1}-1)-(\rho-1)(M+1) \geq 0, \text{ (strictly greater if } \rho \neq 1,\ M \geq 1)$$

so that dividing (8) by $\rho^M[(\rho^{M+1}-1)-(\rho-1)(M+1)]$ does not change the sign of (8),

(9)
$$\frac{lnq}{(\rho^{M+1}-1)^2}\left\{\frac{[\rho^M(\rho-1)(M+1)-(\rho^{M+1}-1)]}{\rho^M[(\rho^{M+1}-1)-(\rho-1)(M+1)]}q^i-\rho q^{N-i}\right\}.$$

Now let

$$\gamma(i)=\frac{[\rho^M(\rho-1)(M+1)-(\rho^{M+1}-1)]}{\rho^M[(\rho^{M+1}-1)-(\rho-1)(M+1)]}.$$

The proofs of the lemmas below follow from induction and are omitted.

LEMMA 1: If $1 \leq i < N/2$ then $\gamma(i) < 1$ and if $N/2 < i \leq N$ then $\gamma(i) > 1$.

LEMMA 2: If $1 \leq i < N/2$ then $q^i - \rho q^{N-i} < 0$ and if $N/2 < i \leq N$ then $q^i - \rho q^{N-i} > 0$.

Since $lnq \leq 0$ this means that the derivative of (7) is positive for $1 \leq i < N/2$ and negative for $N/2 < i \leq N$. We therefore have the following result.

THEOREM 1: Assume that the number of machines, N, is even. Then for any positive integer M and any q, the optimal placement of the buffer is immediately after machine $N/2$.

PROOF: The proof of this theorem is essentially complete. We must only show that (7) is continuous at $i=N/2$. By definition the steady state probability of obtaining a unit from a given cycle when $i=N/2$ is

$$q^N+(q^{N/2}-q^N)\frac{M}{M+1}.$$

When $i \to N/2$ then $\rho \to 1$ so that in (7)

$$\lim_{i \to N/2} \text{ssp(i)}$$

he indeterminate form 0/0. But an application of l'Hospital's Rule [9] shows that

$$\lim_{i \to N/2} \frac{\rho(i)^{M+1} - \rho(i)}{\rho(i)^{M+1} - 1} = \frac{M}{M+1},$$

since

$$\lim_{i \to N/2} \mu(i) = q^{N/2} - q^N$$

continuity is proven.

Q.E.D.

SEQUENTIAL RELAY MODEL AND COMPARISONS

In this section we shall consider the mathematical formulation of a two-machine sequential relay del and make some reasonable comparisons between the sequential relay and serial model per- nance characteristics. Consider a two-machine sequential relay model with a buffer capacity of nits ($M \geq 1$) placed between the two machines. Assume that machine i has reliability q_i and that n machines have identical unit processing times. If X_n is the state of the buffer at the beginning ycle n then

$$P[X_{n+1} = 0 \mid X_n = 0] = p_1$$

$$P[X_{n+1} = 1 \mid X_n = 0] = q_1$$

e, if $X_n = 0$, the second machine necessarily is idle during the nth cycle. This is the only difference the serial model. The transition probabilities from all other states would be identical to the serial el, i.e., Figure 4 with $\alpha_i = \beta_i = q_i$.

The steady state distribution of the buffer size for the sequential relay model is for,

$$0 \leq t \leq M, \; q_1 \neq q_2$$

$$P[X = t] = \begin{cases} \left[\dfrac{q_2 - q_1}{q_2 - q_1 \rho^M} \right] \dfrac{q_1 \rho^{t-1}}{q_2 p_1} & t \geq 1 \\ \dfrac{q_2 - q_1}{q_2 - q_1 \rho^M} & t = 0 \end{cases}$$

e $\rho = q_1 p_2 / q_2 p_1$.

When $q_1 = q_2 = q$

$$P[X = t] = \begin{cases} \dfrac{1-q}{M+1-q} & t = 0 \\ \dfrac{1}{M+1-q} & 1 \leq t \leq M. \end{cases}$$

The long run probability of producing a unit in a given cycle is

$$q_2 \cdot P[X \geq 1]$$

or

$$\frac{q_1 q_2 (1-\rho^M)}{q_2 - q_1 \rho^m} \qquad q_1 \neq q_2$$

and

$$\frac{qM}{M+1-q} \qquad q_1 = q_2 = q.$$

The crucial difference between the sequential relay and serial models is the manner in which t goods flow through the line. For the sequential relay line units flow from machine to buffer to machi whereas in the serial line units flow from machine to machine and buffers are only accessed in t event of machine failure. The question arises as to the comparability of the two models. Consider t case $N=2$ machines. Assume that the reliability of each machine is q for both models, and assu further that the cycle times are identical. This latter assumption implies that the machine processi time for the serial model is twice as fast as that for the sequential relay model. But the sequential re line will require two operators while only one is needed for the serial line. The following theorem p vides bounds for the steady state output of the sequential relay line in terms of the serial line wh $N=2$. Recall from (6) that the steady state probability of producing a unit for the serial model when t buffer capacity is M is

$$q^2 + (q-q^2)\frac{M}{M+1} = \frac{q(M+q)}{M+1}.$$

THEOREM 2: Assume that the cycle times and machine reliabilities, q, for the two-machine se and sequential relay lines are identical and that $M \geq 1$. Then

$$\frac{q(M+q)}{M+1} \geq \frac{qM}{M+1-q} \geq \frac{q(M-1+q)}{M},$$

i.e., the output of the sequential relay line with buffer capacity M is bounded above by a serial line w the same capacity and bounded below by a serial line with capacity $M-1$.

PROOF: Suppose that the first inequality was not true, then

$$(M+1-q)(M+q) < (M+1)M$$

which after simplification yields

$$q < q^2.$$

Similarily suppose that the second inequality was false, i.e.

$$M^2 < (M+1-q)(M-1+q).$$

After some cancellation one obtains

$$0 < -1 + 2q - q^2$$

$$0 < -(1-q)^2.$$

RELATIONSHIP AND COMPARISON WITH OTHER PRODUCTION MODELS

In this section we briefly review the work of Koenigsberg and Buzacott and present certain relaships of their results with those of the present paper. But first we consider the scope and limitations ie serial model of section 3. The serial model of this paper characterizes the operating characteris- of each machine by one parameter q_i, i.e. the distribution of down time is geometric with parameter $1-q_i$. Suppose the distribution of down time cycles is given independently of q_i. As an approxima- one may attempt to model this more general case by defining U_i and D_i as the mean length of a od of up cycles and down cycles respectively and define

$$q_i' = \frac{U_i}{U_i + D_i}$$

$$p_i' = \frac{D_i}{U_i + D_i}.$$

q_i', p_i' can be used in the serial model to obtain an approximation of the steady state output. We consider the implications of this approximation in a simple but illustrative case.

The disturbing feature of using the approximations q_i' and p_i' is that $U_i = D_i = 100$ cycles would uce the same parameters as $U_i = D_i = 1$ cycle. Intuitively one would expect that the latter case d be more desirable in view of the finite capacity buffer of size M. That in general this should be is illustrated by the following example. Consider a two-machine serial production line with a buffer city $M=1$. The first machine has reliability q, and the second machine is characterized by a two- Markov chain as follows:

$$\begin{array}{c} \\ U \\ D \end{array} \begin{array}{c} \begin{array}{cc} U & D \end{array} \\ \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \end{array}$$

robability that the second machine is up (U) during cycle $n+1$ is a_{11} if the second machine was ring cycle n, while the probability is a_{21} if the second machine was down during cycle n. Of course st machine is also described by a two-state Markov chain, but here both rows would be of the form . The steady state distribution for the second machine is of the form

$$(\gamma_1, \gamma_2) = \left(\frac{a_{21}}{a_{21} + a_{12}}, \frac{a_{12}}{a_{21} + a_{12}} \right),$$

while its limiting covariance is

$$\Gamma = \lim\{E(X_n \cdot X_{n+1}) - E(X_n)E(X_{n+1})\} = \gamma_1(a_{11} - \gamma_1)$$

where $X_n = 1$ corresponds to up and $X_n = 0$ corresponds to down. The covariance for the first machi is, of course, 0.

The two-machine serial line can be described via a four-state Markov chain where the state $(1, U$ means that a unit existed in the buffer at the beginning of cycle n and that the second machine was during cycle n:

$$\begin{array}{c} \\ 1U \\ 1D \\ 0U \\ 0D \end{array} \begin{array}{c} \begin{array}{cccc} 1U & 1D & 0U & 0D \end{array} \\ \left\langle \begin{array}{cccc} qa_{11} & qa_{12} & pa_{11} & pa_{12} \\ a_{21} & a_{22} & 0 & 0 \\ 0 & 0 & a_{11} & a_{12} \\ qa_{21} & qa_{22} & pa_{21} & pa_{22} \end{array} \right\rangle \end{array}$$

Note that if the system is in state $(1, D)$ at the beginning of cycle n, then the first machine is blocke during cycle n. Also note that if the system is in state $(0, U)$ at the beginning of cycle N, then at th beginning of cycle $N+1$ there can never be any units in the buffer. This is simply a consequence the definition of the serial line.

The steady state distribution of this four-state Markov chain is $(\pi_1, \pi_2, \pi_3, \pi_4) =$

$$\left(\frac{qa_{21}}{q + a_{21}}, \frac{qa_{22} - q^2(a_{11} - a_{21})}{q + a_{21}}, \frac{qa_{21} + q^2(a_{11} - a_{21})}{q + a_{21}}, \frac{pa_{21}}{q + a_{21}} \right)$$

when (γ_1, γ_2) is assumed equal to (q, p), i.e., both machines have the same steady state reliability Now the steady state probability of obtaining a unit from this two-machine serial system would b $\pi_1 + q\pi_3$ since a unit is always generated when in state $(1, U)$, and if the first machine is up a finishe good unit is obtained when the system is in state $(0, U)$. The expression $\pi_1 + q\pi_3$ can be simplified t

$$\frac{q(a_{21} + a_{12}) + q^2}{(a_{21} + a_{12}) + 1}.$$

Next observe that the covariance, Γ, can be rewritten in the form

$$\Gamma = pq[1 - (a_{21} + a_{12})]$$

hat substituting for $a_{21}+a_{12}$ yields for $\pi_1+q\pi_3$

$$q\left[\frac{pq(1+q)-\Gamma}{2pq-\Gamma}\right].$$

serve now that since $2pq > pq(1+q)$, then $\pi_1+q\pi_3$ is a decreasing function of Γ. Hence the steady e output is larger when Γ is negative, smaller when Γ is positive. When $\Gamma=0$ one obtains $(q+q^2)/2$ ch corresponds to (6) when $M=1$.

This example case indicates that when the covariance between X_n and X_{n+1} is positive, then using approximation q, p for the serial model overestimates the steady state output, while the reverse is : if the variables are negatively correlated. In general, one would expect that positive correlation ne typical case. Now consider an early production line model where the role of the covariance can asily examined.

In [13] Koenigsberg discusses a certain production line which he attributes to Finch. The Koenigs-g-Finch (KF) model is for a production line with two machines and a single buffer stock with finite acity M such as depicted in Figure 4. In the KF model each machine is assumed to be in either one wo states, working or not working. The processing time per unit for machine i is a constant of , $i=1, 2$. The distribution of working time for machine i is assumed to be the continuous exponential dom variable with mean $1/\lambda_i$, and the distribution of down time is also exponentially distributed 1 mean $1/\mu_i$. The problem posed by KF is the determination of the gain in productive output as nction of the buffer size.

In the subsequent derivation given by KF it is assumed that the output of a two-machine line 1 a buffer can differ from the output of a two-machine line with no buffer only when the first machine own. Hence the output of either line, buffer or no buffer, in the case when both machines are king is the same. This is the assumption in the serial production line.

Observe now that λ_i/μ_i is the ratio of the mean down time to the mean up time for stage i. For case $\lambda_1/\mu_1=\lambda_2/\mu_2=U$ and $h_1=h_2=h$, KF show that the ratio of the expected increase in units 1 a buffer of capacity M divided by the expected units with no buffer is

$$U\left\{1-\frac{h(1+U)(\mu_1+\mu_2)}{h(1+U)(\mu_1+\mu_2)+M(\mu_1+\mu_2U)(\mu_2+\mu_1U)}\right\}.$$

we make the assumptions that

(1) both machines are identical, i.e. $\mu_1=\mu_2=\mu$, $\lambda_1=\lambda_2=\lambda$ and $h_1=h_2=h$(units/time)

(2) only one machine is operated at a time during a given cycle, i.e. a single operator runs both machines; this implies an actual production rate of $h/2$. Assume h is unity.

) these assumptions (10) reduces to

$$\left(\frac{\lambda/(\lambda+\mu)}{\mu/(\lambda+\mu)}\right)\left[\frac{M}{M+\dfrac{1}{\lambda+\mu}}\right].$$

Substituting $p' = \lambda/(\lambda+\mu)$, $q' = \mu/(\lambda+\mu)$ which are respectively the ratio of mean down time a mean up time to total time yields

$$\frac{p'}{q'}\left[\frac{M}{M+\frac{1}{\lambda+\mu}}\right] \tag{12}$$

Now we multiply (12) by $(q')^2$ to yield the absolute increase in output due to the buffer capacity

$$q'p'\left[\frac{M}{M+\frac{1}{\lambda+\mu}}\right] \tag{13}$$

Observe now that (13) is identical to the marginal gain for the serial line when $N=2$ and $\lambda+\mu=$ The implication of $\lambda+\mu=1$ can be seen when the exponential assumptions of Koenigsberg are d cretized to obtain a two-state Markov chain with transition matrix

$$\begin{array}{c} \\ U \\ D \end{array}\begin{array}{c} \begin{array}{cc} U & D \end{array} \\ \begin{pmatrix} 1-\lambda\Delta t & \lambda\Delta t \\ \mu\Delta t & 1-\mu\Delta t \end{pmatrix} \end{array}$$

where λ is the breakdown/cycle, μ the repair/cycle, and $\Delta t = 1$ the cycle length. When $\lambda+\mu=1$, th rows of the transition matrix are identical. Otherwise a covariance exists and $\lambda+\mu$ can vary betwe 0 and 2 which means

$$0 \leqslant \frac{M}{M+\frac{1}{\lambda+\mu}} \leqslant \frac{M}{M+\frac{1}{2}} \cdot \tag{14}$$

In later work Buzacott [4, 5, 6, 7] has considered a wide variety of production line models. The papers consider random service times as well as fixed cycle service times. In particular, Buzacott "Automatic Transfer Lines With Buffer Stocks" [4] deals directly with the question of positioning buffer stocks and related issues. A key assumption made by Buzacott [4] is that "If the line has r buffers it will be forced down as soon as any station breaks down." This is the assumption of the seri production line as contrasted with a sequential relay production line.

In [4] Buzacott examines the case where, in addition to a failure rate p_i for the ith machin the repair time for the ith machine is random with mean R_i. For this model Buzacott has develope an approximating technique for estimating the mean throughput of the production line, and has a serted that the throughput is maximized when the buffer is placed, so that the line is separated int two stages of equal reliability, e.g., the middle of the line if all machines have the same failure rat Assuming that the probability of failure of two or more machines in the same cycle is negligible, Buzaco approximates the unbuffered reliability, A_0 for an N machine line with

$$A_0 = \frac{1}{1 + \sum_{i=1}^{N} p_i R_i}.$$

the case when $p_i = p$, $R_i = 1$, this approximation is simply

$$A_0 = \frac{1}{1 + Np}.$$

actual reliability for this case would be q^N where $q = 1 - p$.
Buzacott subsequently defines, A, the reliability of a two-stage (one buffer with capacity M) uction line to be

$$A = A_0 + P_1 h(M)$$

e P_1 is the proportion of time that the first stage is under repair, and $h(M)$ is the proportion of the first stage is under repair that the line is up. According to the approximating procedure if the machine is under repair the second machine is necessarily up. If this simplifying assumption is ped, then the proportion of time the first stage is under repair that the second stage is up would $h(M)$. For the case $p_1 = p_2 = p$, $R_1 = R_2 = 1$, then Buzacott shows that $h(M) = M/M + 1$ so that

$$A = A_0 + p\frac{M}{M+1}$$

$P_1 = p$. Now observe that if $qh(M)$ is used in place of $h(M)$, then

$$A = A_0 + p \cdot q\frac{M}{M+1}$$

$$A = A_0 + (q - q^2)\, M/M + 1.$$

unbuffered reliability is q^2, then this result concurs with Equation (6).

COST/RELIABILITY TRADEOFF FOR THE SERIAL MODEL

In this section we shall derive a sufficient condition for the existence of a buffer in the serial el when there is a possible trade-off of either increasing q or creating a buffer. We shall assume all machines have equal probability of functioning correctly. Let $g(M)$ be the cost of creating a r of size M, where M is a continuous variable, and let $h(q)$ be the cost of maintaining a performance of q for all machines. We shall consider the following maximization problem (MP):

$$\text{(MP)} \qquad \max f(M,q) = q^N + (q^{N/2} - q^N)\frac{M}{M+1}$$

$$\text{subject to: } g(M) + h(q) \leq C$$

$$q \leq 1$$

$$M,\ q \geq 0$$

where C represents the funds available for maintaining the production line. We shall assume that bo g and h are strictly increasing functions, and that when q is near 1 $h(q) > C$. The latter assumption necessary for (MP) to be meaningful, for if $h(1) \leq C$, then one obviously could employ as an optim solution $M = 0$, $q = 1$. Some reasonable h functions could be

$$\text{(i)} \quad h_1(q) = b_1 q^{r_1} \qquad b_1 > 0, r_1 > 0$$

$$\text{(ii)} \quad h_2(q) = b_2/(1-q)^{r_2} \qquad b_2 > 0,\ r_2 > 0$$

where in (i) $b_1 > C$ so that $h_1(1) > C$. Furthermore note that $h_2(q) \to \infty$ as $q \to 1$. We assume that $g($ $+ h(0) < C$; hence Slater's Constraint Qualification is satisfied [15].

Now suppose that $(0, q^*)$ is an optimal solution to (MP), i.e., $M = 0$. Since $f(M, q)$ is a strictl increasing function of q for any fixed $M \geq 0$ it follows that

$$h(q^*) = C$$

or then

$$q^* = h^{-1}(C).$$

For $(0, q^*)$ to be optimal for (MP), $(0, q^*)$ must satisfy the Kuhn-Tucker necessary conditions [11 that $(\lambda_1^*, \lambda_2^*) \geq 0$ exist such that

$$\frac{\partial f}{\partial M}(0, q^*) - \lambda_1^* g'(0) \leq 0 \tag{15}$$

$$\frac{\partial f}{\partial q}(0, q^*) - \lambda_1^* h'(q^*) - \lambda_2^* \leq 0. \tag{16}$$

Observe that based upon our assumptions concerning h, it must follow that $q^* < 1$, and this in turn implies that $\lambda_2^* = 0$. Furthermore note that (16) will be satisfied as a strict equality since $q^* > 0$ Hence from (16) one would obtain

$$\lambda_1^* = \frac{\frac{\partial f}{\partial q}(0, q^*)}{h'(q^*)},$$

d substituting (17) into (15) yields

$$\frac{\partial f}{\partial M}(0, q^*) \leq \frac{g'(0)}{h'(q^*)} \cdot \frac{\partial f}{\partial q}(0, q^*).$$

er inserting the appropriate quantities into (18) one obtains

$$q^{*N/2} - q^{*N} \leq \frac{g'(0)}{h'(q^*)} N q^{*(N-1)},$$

ich after cancelling a $q^{*N/2}$ and rearranging yields

$$q^{*N/2}\left[\frac{g'(0)N}{h'(q^*)\cdot q^*} + 1\right] \geq 1.$$

THEOREM 3: A sufficient condition for the existence of a buffer (immediately after machine) is that

$$q^{*N/2}\left[\frac{g'(0)N}{h'(q^*)\cdot q^*} + 1\right] < 1$$

re q^* is the maximum attainable reliability with no buffer, i.e., $h(q^*) = C$.

PROOF: If the hypothesis of the theorem is satisfied, then $(0, q^*)$ could not be optimal since (19) olated. Hence the optimal solution must be of the form (M, q) where $M > 0$.

The optimal solution to (MP) can be obtained in a straightforward manner that is especially well ed to the case when M must be integer. At optimality the constraint $g(M) + h(q) \leq C$ is obviously sfied as an equality so that q can be expressed as an explicit function of M, i.e., $q = h^{-1}[C - g(M)]$. ce, the objective function can be formulated as a function of M alone and the optimal solution can btained by maximizing over the set of feasible values for M.

In summary we should analyze the relationship (19) more closely. Note that (19) would be violated if

(i) the marginal cost of buffer capacity at zero is small, i.e., $g'(0)$ near zero.

(ii) the marginal cost of increased reliability at q^* is high, i.e., $h'(q^*)$ very large.

(iii) the number of machines in the line is large.

that (iii) holds since

$$q^{*N/2}\left[\frac{g'(0)N}{h'(q^*)q^*} + 1\right] \to 0$$

$\to \infty$, for $Nq^{N/2} \to 0$ since $q < 1$, see [15].

ACKNOWLEDGMENT

The authors wish to thank Professor Matt Rosenshine of the Industrial Engineering Departmen the Pennsylvania State University, for several enlightening discussions concerning this paper. I particular, Professor Rosenshine brought to our attention the possible role of the covariance functio

REFERENCES

[1] Bartle, R. G., The Elements of Real Analysis (John Wiley, New York, 1964).

[2] Burke, Paul J., "The Output of a Queuing System," Operations Research *4*, No. 6 (1956).

[3] Buxey, G. M., N. D. Slack and Ray Wild, "Production Flow Line Systems Design—A Review AIIE Transactions (March 1973).

[4] Buzacott, J. A., "Automatic Transfer Lines with Buffer Stocks," International J. Production Re *5*, No. 3 (1967).

[5] Buzacott, J. A., "The Effect of Station Breakdowns and Random Processing Times on the Capaci of Flow Lines with In-Process Storage," AIIE Transactions, *4*, (1972).

[6] Buzacott, J. A., "The Role of Inventory Banks in Flow-Line Production Systems," The Inte national J. Production Res. *9*, No. 4 (1971).

[7] Buzacott, J. A., "Prediction of the Efficiency of Production Systems Without Internal Storage, International J. Production Res. *6*, No. 3 (1968).

[8] Clarke, A. Bruce and Ralph L. Disney, Probability and Random Processes for Engineers an Scientists (John Wiley and Sons, New York, 1970).

[9] Courant, R., *Differential and Integral Calculus*, Vol. *1*, Interscience (John Wiley, New York, 1964

[10] Freeman, Michael C., "The Effects of Breakdowns and Interstage Storage on Production Lin Capacity," J. Indust. Engin. *15*, No. 4 (1964).

[11] Hillier, Frederick S. and Ronald W. Boling, "Finite Queues in Series with Exponential or Erlan Service Times—A Numerical Approach," Oper. Res, *15*, No. 2 (1967).

[12] Hunt, Gordon C., "Sequential Arrays of Waiting Lines," Oper. Res., *4*, No. 6 (1956).

[13] Koenigsberg, E., "Production Lines and Internal Storage—A Review," Man. Sci. *5*, No. 4 (1959)

[14] Kuhn, H. W. and A. W. Tucker, "Nonlinear Programming," in *Proceedings of the Second Sym posium on Mathematical Statistics and Probability*, J. Neyman (ed.) (University of Californi Press, Berkeley, Calif., 1951).

[15] Mangasarian, O. L., Nonlinear Programming (McGraw-Hill, New York, 1969).

DETERMINING ADJACENT VERTICES ON ASSIGNMENT POLYTOPES

Patrick G. McKeown

State University of New York at Albany
School of Business
Albany, New York

ABSTRACT

To rank the solutions to the assignment problem using an extreme point method, it is necessary to be able to find all extreme points which are adjacent to a given extreme solution. Recent work has shown a procedure for determining adjacent vertices on transportation polytopes using a modification of the Chernikova Algorithm. We present here a proccdure for assignment polytopes which is a simplification of the more general procedure for transportation polytopes and which also allows for implicit enumeration of adjacent vertices.

Introduction

In a recent paper, McKeown and Rubin [4] demonstrated that a modified and simplified version the Chernikova Algorithm [3] could be used to generate adjacent vertices on transportation polytopes. his modification for transportation problems uses the special structure of the constraint set to reduce e amount of checking required before two columns can be combined to form an edge leading to a w vertex.

However, this procedure procludes dropping edges for purely cost considerations, i.e., even if an ge could be shown not to satisfy some upper bound, it must be retained for checking later edges. is would become very unwieldy if one wished to rank the vertices of the assignment problem in ndecreasing order according to objective value. In such cases, there will be $\sum_{i=0}^{n-2} \binom{n}{i} (n-i-1)!$ edges nerated [1]. Such a ranking procedure would be used in two situations. First, one might wish to termine the best solution other than the optimal that would satisfy some sort of secondary con- aints, possibly behavioral. Secondly, it is well known that various problems with assignment nstraints can be solved by ranking the vertices of a related assignment problem in a manner similar that presented by Murty [6] for the linear fixed charge problem. Examples of such problems are the adratic assignment problem [2], the traveling salesman problem [8], and the bottleneck assignment oblem.

We will present here a theorem which shows that, for the assignment polytope, we may avoid eping any columns of the Chernikova tableau which are unwanted from a cost standpoint. In so doing, are able to implicitly enumerate the adjacent vertices rather than actually finding all of them. This per will present a theoretical framework for such a procedure; a later paper will present the pro- lure and computational results for various applications [5]. Before doing so, we will present a discus-

sion and paraphrase of the Chernikova algorithm for transportation polytopes. For a more detailed description, see [4].

II. THE CHERNIKOVA ALGORITHM FOR TRANSPORTATION POLYTOPES

The assignment problem may be formulated as follows:

$$\text{Min} \sum_{i=1}^{n} \sum_{j=1}^{n} C_{ij} X_{ij}$$

$$\text{S.t.} \sum_{i=1}^{n} X_{ij} = 1 \qquad j = 1, \ldots, n, \tag{1}$$

$$\sum_{j=1}^{n} X_{ij} = 1, \qquad i = 1, \ldots, n, \qquad (P) \tag{2}$$

$$X_{ij} \geq 0. \tag{3}$$

Since P is a special case of the transportation problem, the solutions will be all integer. In addition it is important to note that P is degenerate to degree $n-1$. This degeneracy has caused difficulties in finding vertices adjacent to any vertex of the polyhedron defined by the constraint set of P.

The Chernikova algorithm finds adjacent vertices by determining *edges* that connect the present vertex to the adjacent vertices. It should be remembered that this is *geometric adjacency* as compared to the *algebraic adjacency* associated with a simplex pivot. For a more detailed discussion of these concepts, see [4]. In either case, the edge leading to an adjacent extreme solution may be represented in the context of the transportation or assignment tableau by a *loop*, i.e., a nondecomposable simple connected path that connects one or more nonbasic cells to a subset of the basic cells. The *edge directions* are found by starting with a nonbasic cell, assigning a -1 and proceeding around the loop alternately assigning $+1$, -1, etc. If a $+1$ falls in a degenerate basic cell, then this edge leads to an algebraically adjacent vertex but not to a geometrically adjacent vertex, i.e., we find another basic representation of the present vertex. If no $+1$'s fall in degenerate basic cells, then we have a geometrically adjacent vertex. It is this latter condition that the Chernikova algorithm generates. We will now describe that the algorithm for the more general case of the transportation problem.

Recall that in Tucker's tableau for some solution (X_B, X_N),

$$X_B = B^{-1}b + B^{-1}N(-X_N) \geq 0$$

and

$$X_N = 0 + (-I)\,(-X_N) \geq 0.$$

Then we may partition X_B into degenerate and nondegenerate variables as follows:

$$X_B = (X_Q, X_D)$$

where

$$X_Q = q + Q(-D_N) \geq 0,$$

$$X_D = 0 + D(-X_N) \geq 0.$$

$$X_N = 0 + (-I)\ (-X_N) \geq 0, \text{ and } q > 0.$$

Chernikova algorithm finds vertices adjacent to the present solution by determining the edges of one $C = \{X_N | -DX_N \geq 0,\ X_N \geq 0\}$. To do this, first consider the matrix $\binom{-D}{-I}$. Chernikova's algorithm s a series of transformations of this matrix which generate all edges corresponding to adjacent ces. At any stage of the process, we denote the old matrix by $Y = \binom{U}{L}$, and the new matrix being rated is denoted $\bar{Y}$. For transportation polytopes, all elements of Y are $+1$, -1, or 0 always [4]. ır case, the simplex tableau less the nondegenerate rows can serve as the initial Y matrix for nikova's algorithm. If the final tableau in the algorithm is $\binom{U}{L}$, then the rows corresponding to the egenerate basic variables have entries $-QL$. For the transportation tableau, refer to columns of Chernikova tableau by two systems: a single subscript k indicates that we refer to the kth column e tableau; a double subscript refers to the cell in the assignment tableau associated with this nn. For initial columns k or ij (before Chernikova's procedure is begun) let $R(k) = \{i\}$, $C(k) = \{j\}$, $(k) = \{ij\}$. Note that at this point, $I(k)$ contains the subscripts of the $(n-1)^2$ nonbasic variables. ow replace the matrix L with the index sets R, C, and I. Using these modifications, we may state llowing:

ALGORITHM

).1. If all elements of U are nonpositive, then the algorithm terminates with the edges contained).

1) Choose the first row of U, say row r, with at least one positive element.

2) Let $R = \{j | y_{rj} < 0\}$. Let $v = |R|$. Then the first v columns of the new matrix, $\bar{Y}$, are all the y_j for $j \epsilon R$, where y_j denotes the jth column of Y.

3) Let $S = \{(s,t) | y_{rs} y_{rt} < 0,\ s < t\}$, and I_0 be the index set of all nonpositive rows of U. For each $(s,t) \epsilon S$, let $I_1(s,t) = \{i \epsilon I_0 | y_{is} = y_{it} = 0\}$. We now use some of the elements of S to create additional columns of $\bar{Y}$.

(a) If $R(s) \cap R(t) \neq 0$ or $C(s) \cap C(t) \neq 0$, don't combine columns s and t. Get the next element of S, and go to (a).

(b) For each column $v \neq s,t$, proceed as follows

(i) If $I(v) \not\subseteq I(s) \cup I(t)$, go to the next v.

(ii) If $I(v) \subseteq I(s) \cup I(t)$, check to see if $y_{iv} = 0$ for all $i \epsilon I_1(s,t)$ or if $I_1(s,t) = 0$. If so, don't combine s and t, get the next element of S, and go to (a).

(iii) If all $v \neq s,t$ have been checked and none prevented the combination of s and t, then combine columns s and t to form a new column of $\bar{Y}$. This combination is made as follows:

$$U_{s+t} = U_s + U_t,$$

$$R(s+t) = R(s) \cup R(t),$$

$$C(s+t) = C(s) \cup C(t),$$

$$I(s+t) = I(s) \cup I(t).$$

Get the next element of S, and go to (a).

(4) When all pairs in S have been examined, and the additional columns (if any) have been ad we say that row r has been processed. Now let Y denote the matrix $\bar{Y}$ produced in proces row r, and return to step 0.1.

Costs for adjacent vertices may be computed by adding a dual variable row to the Y matrix. set of checks in step (3)(b) above preclude dropping columns that can be shown not to satisfy s present cost requirement. This could happen if we did not want to generate edges which corresp to assignments with value worse than some upper bound. The next section will show that for such ca we need not keep these unwanted edges.

III. MODIFICATIONS OF THE CHERNIKOVA ALGORITHM

We may note from the above discussion of the Chernikova algorithm that there is a considera amount of checking necessary to generate all adjacent vertices. It is also evident that columns n not be deleted due to cost considerations since they must be kept for the checking procedure. For assignment polytope, the following theorem will show that we need not do the (3)(b) checking, i only the columns s and t need be checked before they are combined.

THEOREM: For the assignment problem, consider the diagonal assignment (i.e., $\bar{X}_{ij}=1$, $i=1, \ldots, n$) with the $n-1$ degenerate assignments all chosen to be in the first row of the assignment table. If for two columns of the Chernikova tableau, say s and t, we have $y_{rs}y_{rt}<0$, $R(s)\cap R(t)=\emptyset$, a $C(s)\cap C(t)=\emptyset$, then columns s and t may be combined. We note that since we can renumber variables, any result about the diagonal assignment is valid for all assignments.

Before proving this theorem, we will state two lemmas which are essential to the proof of theorem:

First recall in the following lemma that the U matrix contains the degenerate elements of Assignment Tableau. We will then adopt the notation that U_k refers to the kth column of U, and is the (i,k) element of that matrix. We will also use $|U_k|$ to mean the number of elements in U_k.

LEMMA 1: under the conditions given the theorem, for some columns of the Chernikova table we have

(i) $|U_s|=0$, 1, or 2; moreover, for some rows i and j we have

(ii) $u_{is}=-1$, if $|U_s|=1$, and

(iii) $u_{is}u_{js}=-1$, if $|U_s|=2$.

PROOF: From the notion of a loop, we know that no row or column of the tableau can have m than two elements in a loop. Moreover, if there are two elements in the same row or column, they m have opposite edge directions. Since U_s contains the edge directions in the degenerate cells for so loop, (i) and (iii) follow directly. For (ii), assume that $|U_s|=1$, and for any i, $u_{is}=1$. This implies that t ith degenerate cell is connected to cell (1,1) with edge sign -1 and that cell (1, 1) is connected t nonbasic cell $(k, 1)$ with edge sign $+1$. But, this is a contradiction of the loop rules and u_{is} must ha edge sign -1. Q.E.

LEMMA 2: Under the conditions given in the theorem, for any column s of the Chernikova table

(i) $u_{js}=+1 \Rightarrow (i, j+1)\epsilon I(s)$ for some i, and

(ii) $u_{is}=-1 \Rightarrow (i+1, j)\epsilon I(s)$ for some j.

PROOF (i): We know that $u_{js}=+1$ implies that for some nonbasic cell s, the loop joining s a subset of the basic cells has an edge direction of $+1$ in the jth degenerate cell. Note also that the

ierate cell is also the $(j+1)$st cell in the first row. By the concept of a loop, the cell $(1, j+1)$ must nnected to some cell $(i, j+1)$ with edge sign -1. Let's assume that $(i, j+1)$ is basic. Since we are ng with the diagonal assignment, $i=j+1$, and there are no other nondegenerate basic cells in ow. If there is a nonbasic cell in this row, it must have edge sign -1 and the basic cell $(j+1, j+1)$ have edge sign $+1$. But this contradicts the above assumption. Hence the cell $(i, j+1)$ must be sic and a member of $I(s)$.

he proof for the second case is similar. Q.E.D.

ROOF OF THEOREM: If s and t are the columns to be combined where $R(s) \cap R(t) = \emptyset$ and $C(t) = \emptyset$, then by Lemma 1 there are three possible cases to examine. These are:

(i) $u_{rs} = -1$ and $\exists j$ such that $u_{js} = +1$
$u_{rt} = +1$ and $\exists i$ such that $u_{it} = -1$ for $i \neq j$.
ii) same as (i) except $i = j$.
iii) $u_{rs} = -1$
$u_{rt} = +1$ and $\exists i$ such that $u_{is} = -1$.

ill show that if s and t are the edges, then so is $(s+t)$. Note that if $r=1$, we can always combine t. Hence, we may assume that, for $r \neq 1$, s and t are edges, and attempt to prove that $s+t$ is also ge. We will now proceed to prove case (i) above:

ASE I: If $u_{rs} = -1$ and $u_{js} = +1$, then there exists for $j = j_l$ a loop joining the elements $(1, r+1)$, $+1)$, a subset of the elements of the diagonal assignment and the elements of I, i.e., $\{(\underset{(-)}{1}, r+1), (\underset{(+)}{r}, r+1), (r+1, \underset{(-)}{i_1}), \ldots, (\underset{(+)}{i_l}, i_l), (i_l, \underset{(-)}{j_l}+1), (1, \underset{(+)}{j_l}+1)\}$ with edge signs as shown. In particu- Lemma 2 we know that (i_l, j_l+1) and $(r+1, i_1)$ are in this loop. Now if $u_{rt} = +1$ and $u_{it} = -1$, n a similar fashion we will have the following loop elements and edge signs for $i = i_k$: $\{(1, \underset{(+)}{r}+1), (\underset{(+)}{j_1}, j_1), \ldots, (\underset{(+)}{j_k}, j_k), (i_k+1, \underset{(-)}{j_k}), (i_k+1, \underset{(+)}{i_k}+1), (1, \underset{(-)}{i_k}+1)\}$. If we now compute the f s and t, we have: $\{(1, \underset{(+)}{j_l}+1), (i_l, \underset{(-)}{j_l}+1), (\underset{(+)}{i_l}, i_l), \ldots (r+1, \underset{(-)}{i_1}), (r+1, \underset{(+)}{r}+1), (\underset{(-)}{j_1}, r+1), \ldots (\underset{(+)}{j_k}, j_k), (i_k+1, \underset{(-)}{j_k}), (i_k+1, \underset{(+)}{i_k}+1), (1, \underset{(-)}{i_k}+1)\}$. This is a nondecomposable loop repre- g the sum of s and t; hence it corresponds to an edge. Since a degenerate element, $(1, j_l+1)$, positive edge sign, this column corresponds to an alternate basic representation of the diagonal ments, i.e., an edge of zero length.

ases (ii) and (iii) follow in a similar manner except that they correspond to vertices other than the al assignment. It is only necessary to show that these new vertices are adjacent to the diagonal ment. This may be shown using the result proved by Murty [7] that all assignments adjacent diagonal assignment are either traveling salesman tours or subtours with selftours. Thus in cases l (iii) we need only show that the resulting assignments correspond to one of these two possibili- his may be done by looking at the result in case (iii) of combining s and t, i.e. $\{(\underset{+}{1}, 1), (\underset{-}{i_l}, 1), \ldots, (r+1, \underset{-}{i_1}), (r+1, \underset{+}{r}+1), (\underset{-}{j_1}, r+1), (\underset{+}{j_1}, j_1), \ldots, (\underset{+}{j_k}, j_k), (i_k+1, \underset{-}{j_k}), (i_k+1, \underset{+}{i_k}+1), (\underset{-}{1}, 1)\}$. In this case, the new assignment will contain those cells with an edge sign of $(-)$ and any al elements which are not contained in this loop, i.e., $\{(i_l, 1), (i_{l-1}, i_l), \ldots, (r+1, i_1),$

$(j_1, r+1), \ldots (i_K+1, j_K), (1, i_K+1)\}$, which is a tour (or subtour with selftours). Hence $(s+t)$ responds to an adjacent assignment. The reasoning for case (ii) is the same. Q.I

From an intuitive standpoint, what we have shown is that if s and t satisfy the conditions of theorem, then the loops corresponding to the columns s and t may always be combined to form a nondecomposable loop.

A numerical example of this revised algorithm may be found by referring to [4]. The exam there is an assignment problem that has diagonal assignments but has degenerate cells not all in first row. If one were to rearrange the degenerate cells to satisfy the theorem, that example may ea be reworked without the zero checks present in that algorithm.

REFERENCES

[1] Balinski, M. L. and Andrew Russakoff, "On The Assignment Polytope," SIAM Review *16*, 516–(1974).

[2] Cabot, A. V. and R. L. Francis, "Solving Certain Nonconvex Quadratic Minimization Probl by Ranking the Extreme Points," Oper. Res. *18*, 82–86 (1970).

[3] Chernikova, N. V., "Algorithm for Finding a General Formula for the Nonnegative Solutions System of Linear Inequalities," U.S.S.R. Computational Mathematics and Mathematical Phys *5*, 228–233 (1965).

[4] McKeown, P. G. and D. S. Rubin, "Adjacent Vertices on Transportation Polytopes," Nav. F Log. Quart. *22*, 365–374 (1975).

[5] McKeown, P. G. and Donald Bishko, "Using the Chernikova Algorithm to Rank Assignment in preparation.

[6] Murty, K. G., "Solving the Fixed Charge Problem by Ranking the Extreme Points," Oper. R *16*, 268–279 (1968).

[7] Murty, K. G., "On the Tours of a Traveling Salesman," SIAM Journal on Control 7, 122–131 (196

[8] Murty, K. G., Caroline Karel and John D. C. Little, "The Traveling Salesman Problem: Solut by a Method of Ranking Assignments," Unpublished, Case Institute of Technology (196

IMPROVED CONVERGENCE RATE RESULTS FOR A CLASS OF EXPONENTIAL PENALTY FUNCTIONS

Frederic H. Murphy

Federal Energy Administration
Washington, D.C.

ABSTRACT

An improved theoretical rate of convergence is shown for a member of the class of exponential penalty function algorithms. We show that the algorithm has a superlinear convergence rate.

RODUCTION

An exponential penalty function algorithm is presented in [9] for the following nonlinear ramming *NLP*:

$$\underset{x \in E^n}{\text{minimize}}\ f(x)$$

ct to

$$g_i(x) \leq 0 \qquad \text{for } i = 1, \ldots, m$$

an optimal solution x^*. The penalty function is

$$F_k(x) = f(x) - \sum_{i=1}^{m} \frac{1}{s_k} \exp\left[r_k g_i(x)\right]$$

$r_k \geq s_k \geq 1$ and $r_k \to \infty$. This class of penalty functions is contained in the class of penalty ons discussed in Evans and Gould [3]. And O'Neill and Widhelm [12] successfully used a variant is penalty function as the Lagrangian of a modified version of Dantzig-Wolfe generalized amming.

HOD

variant of (3) is considered where we can analyze its asymptotic convergence properties. It is that in the limit it has the convergence properties of method of multiplier algorithms [13] with . This version of (3) also meets one of the important concerns with method of multiplier type thms, that is, it is everywhere second order differentiable if $f(.)$, and $g_1(.), \ldots, g_m(.)$ are. Our hm is:

Step 1) Minimize

$$f(x) + \sum_{i=1}^{m} \frac{\lambda_i^k}{r_k} \exp\left[r_k g_i(x)\right]$$

with the minimum at x_k.

(*Step 2*) Set

$$\gamma_i^k = \lambda_i^k \exp\left[r_k g_i(x_k)\right], \tag{5}$$

then let

$$\lambda_i^{k+1} = \min\left[U_k, \max\left[L_{k+1}, \gamma_i^k\right]\right] \tag{6}$$

where $U_k > L_{k+1} > 0$, $U_k \to \infty$ and $L_{k+1} \to 0$, and $r_k L_{k+1} \to \infty$, and $U_k/r_k \to 0$ as $k \to \infty$.

CONVERGENCE CONSIDERATIONS

Since this penalty function is a member of the class of exponential penalty functions as describ by Évans and Gould [3] we need not prove convergence here. To establish our convergence rate resu we make the following assumptions:

1. NLP is a convex programming problem.
2. Slater's constraint qualification [7] is satisfied.
3. Either $f(x)$ or one of the binding constraints has a positive definite Hessian at the optin solution.
4. The feasible region is compact.
5. Strict complementarity holds, that is, $g_i(x^*) = 0$ implies $\lambda_i^* > 0$ where λ_i^* is the optimal Lagran multiplier for constraint i, $i = 1, \ldots, m$.
6. $\lambda_1^*, \ldots, \lambda_m^*$ are unique, x^* is a unique optimal solution.
7. $\{\nabla g_i(x^*) \mid g_i(x^*) = 0 \; i = 1, \ldots, m\}$ are linearly independent.

Assumption 4 and continuity as implied by assumption 1 guarantee convergence of the algorithm. [10] it was shown that under assumptions 1 and 2, $\gamma_1^k, \ldots, \gamma_m^k$ are uniformly bounded. As this is standard proof (see [4]), it is not repeated here. In constructing a dual function for a modified versi of NLP we need assumptions 3 and 7 to be able to invert a matrix in the dual function.

LEMMA 1: Let $I \subset \{1, \ldots, m\}$ be the index set of the binding constraints at the optimal solutio With assumptions 1, 2, 5, and 6,

$$\lim_{k \to \infty} r_k g_i(x_k) = 0 \qquad \text{for } i \in I. \tag{7}$$

PROOF: For $i \in I$, there is a K such that for $k \geq K$

$$\lambda_i^{k+1} = \min\left[U_k, \max\left[L_{k+1}, \gamma_i^k\right]\right] = \gamma_i^k. \tag{8}$$

Since $\gamma_i^k \to \lambda_i^*$ as $k \to \infty$ by the uniqueness of λ_i^*, $\lambda_i^k \to \lambda_i^*$ as $k \to \infty$. That is,

$$\begin{aligned} \lambda_i^{k+1} - \lambda_i^k &= \lambda_i^k \exp\left[r_k g_i(x_k)\right] - \lambda_i^k \\ &= \lambda_i^k \left[\exp\left[r_k g_i(x_k)\right] - 1\right]. \end{aligned} \tag{9}$$

Since $\lambda_i^{k+1} - \lambda_i^k \to 0$ as $k \to \infty$ and $\lambda_i^* > 0$,

$$\exp\,[r_k g_i(x_k)] \to 1 \text{ as } k \to \infty$$

$$r_k g_i(x_k) \to 0 \text{ as } k \to \infty. \qquad \square$$

ke the algorithms that fit into the classification of methods of multipliers [5], [8], [11], [13], e no trial Lagrange Multiplier is ever zero, with this algorithm we bear the computational burden arrying along nonbinding constraints within the unconstrained minimization. To deal with this blem we have to analyze the behavior of the nonbinding constraints.

LEMMA 2: There exists a K such that for $k \geq K$, and $i \notin I$, $\lambda_i^k = L_k$.

PROOF: For $i \notin I$, since $g_i(x^*) < 0$ there exists an $\epsilon > 0$ and a k such that for $k \geq K$

$$g_i(x_{k-1}) \leq -\epsilon.$$

t is,

$$\exp\,[r_{k-1} g_i(x_{k-1})] \leq \exp\,[-r_{k-1}\epsilon].$$

sequently,

$$\gamma_i^{k-1} \leq \lambda_i^{k-1} \exp\,[-r_{k-1}\epsilon] \leq \frac{U_{k-1}}{r_{k-1}}\, r_{k-1} \exp\,[-r_{k-1}\epsilon].$$

$U_{k-1}/r_{k-1} \to 0$ and $r_{k-1} \exp\,[-r_{k-1}\epsilon] \to 0$, there exists a K' such that the right hand side of (13) s than L_k because $L_k r_{k-1} \to \infty$ as $k \to \infty$. Choose $k \geq \max(K, K')$ and the lemma holds. $\square$

Without loss of generality, let $I = \{1, \ldots, p\}$. We may then formulate a set of new nonlinear ams NLP1:

$$\underset{x \in E^n}{\text{minimize}}\ f_0(r_k, x) = f(x) + \sum_{i=p+1}^{m} L_k/r_k \exp\,[r_k g_i(x)]$$

$$\text{subject to } g_i(x) = 0 \qquad \text{for } i = 1, \ldots, p.$$

ow construct a penalty function for each nonlinear program in *NLP*1. We use the same penalty ion as before with the r_k used in the penalty function the same as that used in the objective func- n 14). Letting $\hat{x}_k$ be the optimal solution to this penalty function we see that $\hat{x}_k \to x^*$ as $k \to \infty$; large enough by Lemma 2 $x_k = \hat{x}_k$; and for r_k large enough (14) is arbitrarily close to (1) over the le region defined by (2). We can now apply the duality results in Luenberger [6, p. 321]. We the dual function ϕ of an element of NLP1 for a given value of r_k as:

$$\phi(\lambda) = \underset{x \in E^n}{\text{minimum}} \left\{ f(x) + \sum_{i=p+1}^{m} [L_k/r_k] \exp [r_k g_i(x)] \right.$$

$$\left. + \sum_{i=1}^{p} [\lambda_i/r_k] [\exp [r_k g_i(x)] - 1] \right\} \tag{16}$$

$$= \underset{x \in E^n}{\text{minimum}} \left\{ f_o(r_k, x) + \sum_{i=1}^{p} [\lambda_i/r_k] [\exp [r_k g_i(x)] - 1] \right\}$$

where $\lambda = (\lambda_1, \ldots, \lambda_p)$.

The function ϕ is convex since we assumed $f(x), g_1(x), \ldots, g_m(x)$ are convex.

Now by [6, p. 321] with $k \geq K$ as defined in Lemma 2

$$\nabla\phi(\lambda^k) = \begin{bmatrix} 1/r_k[\exp [r_k g_1(x_k)] - 1] \\ \cdot \\ \cdot \\ \cdot \\ 1/r_k [\exp [r_k g_p(x_k)] - 1] \end{bmatrix}. \tag{17}$$

$$\text{Let } \nabla g(x) = \begin{bmatrix} \nabla g_1(x) \\ \cdot \\ \cdot \\ \cdot \\ \nabla g_p(x) \end{bmatrix}, \tag{18}$$

$$\text{let } \Lambda(\lambda) = \begin{bmatrix} \lambda_1 & \cdot & \cdot & \cdot & 0 \\ & \cdot & & & \\ & & \cdot & & \\ & & & \cdot & \\ 0 & \cdot & \cdot & \cdot & \lambda_p \end{bmatrix}, \tag{19}$$

$$\text{let } H(r_k, x) = \begin{bmatrix} \exp [r_k g_1(x)] & \ldots & 0 \\ \cdot & & \\ & \cdot & \\ & & \cdot \\ 0 & \ldots & \exp [r_k g_p(x)] \end{bmatrix}, \tag{20}$$

let $F_o(r_k, x)$ be the Hessian of $f_o(r_k, x)$ and let $G_i(x)$ be the Hessian of $g_i(x)$ for $i = 1, \ldots, p$.

From this we can construct the Hessian of $\phi(\lambda^k)$:

$$H(r_k, x_k)\nabla g(x_k) \left[F_o(r_k, x_k) + \sum_{i=1}^{p} \lambda_i^k \exp [r_k g_i(x_k)] G_i(x_k) \right. \tag{21}$$

$$\left. + r_k \nabla g(x_k)' \Lambda(\lambda^k) H(r_k, x_k) \nabla g_i(x_k) \right]^{-1} \nabla g_i(x_k)' H(r_k, x_k)$$

inverse of the matrix in the brackets exists since this matrix is positive definite. That comes
ıt since in assumption 3 we assumed positive definiteness of the Hessian of one of the functions
, $g_1(x)$, . . ., $g_p(x)$ and all other terms in this matrix are positive semidefinite. Now (21) can be
ten as

$$M_k = H(r_k, x_k)\nabla g(x_k)\left[F_o(r_k, x_k) + \sum_{i=1}^{p} \lambda_i^k \exp\left[r_k g_i(x_k)\right] G_i(x_k) \right.$$

$$\left. + r_k \nabla g(x_k)' \Lambda(\lambda^{k+1}) \nabla g_i(x_k) \right]^{-1} \nabla g_i(x_k)' H(r_k, x_k)'.$$

nalyze (22) we now prove a simple extension of a lemma in Mangasarian [8].

LEMMA 3: Let $C(\alpha) = B(A(\alpha) + B'Q(\alpha)B)^{-1}B'$ where B is a given $m \times n$ matrix of rank m,) is an $n \times n$ matrix function on R, $Q(\alpha)$ is a differentiable $m \times m$ matrix function on R, $A(\alpha) +$ $(\alpha)B$ is positive definite for $\alpha \geq \bar{\alpha}$ for some $\bar{\alpha}$, and for every $\epsilon > 0$ there is an α_ϵ such that for α_ϵ, $|A(\alpha) - A| < \epsilon$ where $A + B'Q(\alpha)B$ is positive definite for $\alpha \geq \bar{\alpha}$. Then there exists a constant ix D such that for every $\delta > 0$ there is an α_δ with

$$|C(\alpha) - (Q(\alpha) + D)^{-1}| < \delta$$

$\geq \alpha_\delta$.

PROOF: Define $C_1(\alpha) = B(A + B'Q(\alpha)B^{-1})B'$. By the continuity of the inverse of a matrix, very δ there exists an ϵ_δ such that

$$|C(\alpha) - C_1(\alpha)| < \delta$$

$A(x) - A| < \epsilon_\delta$. This means there exists an α_δ such that (24) holds for $\alpha \geq \alpha_\delta$.

The formula for differentiating the inverse of a matrix is

$$\frac{d(C_1(\alpha)^{-1})}{d\alpha} = -C_1(\alpha)^{-1} \frac{dC_1(\alpha)}{d\alpha} C_1(\alpha)^{-1}.$$

entiating $C_1(\alpha)$ we have

$$\frac{dC_1(\alpha)}{d\alpha} = B \frac{d(A + B'Q(\alpha)B)^{-1}}{d\alpha} B'$$

$$= -B(A + B'Q(\alpha)B)^{-1} \frac{d(A + B'Q(\alpha)B)}{d\alpha} (A + B'Q(\alpha)B)^{-1}B'$$

$$= -C_1(\alpha) \frac{dQ(\alpha)}{d\alpha} C_1(\alpha).$$

ence

$$\frac{d(C_1(\alpha)^{-1})}{d\alpha} = \frac{dQ(\alpha)}{d\alpha} \text{ and } C_1(\alpha)^{-1} = Q(\alpha) + D.$$

That is,

$$C_1(\alpha) = (Q(\alpha) + D)^{-1}. \tag{27}$$

Substituting (27) into (24) we have (23).

For r_k sufficiently large there is an $\epsilon < 0$ where $g_i(x_k) < \epsilon$ for $i = p+1, \ldots, n$. Letting A be t Hessian of $f(x)$ at x_k and $A(r_k)$ be the Hessian of $f_o(r_k, x)$ at x_k for $r_k = \alpha$, we know that $A(r_k) \to$ as $r_k \to \infty$. Let $Q(r_k) = r_k \Lambda(\lambda^{k+1})$ and let $B = \nabla g(x_k)$. Remembering that $H(r_k, x_k) \to I$ as $r_k \to$ we have the following theorem:

THEOREM 1: If NLP satisfies 1, . . ., 7, then the limit of $r_k M_k$ as $k \to \infty$ is $\Lambda(\lambda^*)^{-1}$.

Also, we know from (9) for $i = 1, \ldots, p$

$$\lambda_i^{k+1} - \lambda_i^k = \lambda_i^k [\exp [r_k g_i(x_k)] - 1]. \tag{28}$$

Consequently,

$$\lambda^{k+1} = r_k \lambda^k \Lambda(\lambda^k) \nabla \varphi(\lambda^k). \tag{29}$$

Since $r_k M_k \to \Lambda^{-1}(\lambda^k)$ as $k \to \infty$, we see that, for large r, $\Lambda(\lambda^k)$ approximates the inverse of t Hessian of $\varphi(\lambda)$.

This is important for the following reason. The ith component of the gradient of $\varphi(\lambda^k)$ is

$$\nabla_i \varphi(\lambda^k) = 1/r_k [\exp [r_k g_i(x_k)] - 1] \qquad \text{for } i = 1, \ldots, p. \tag{30}$$

The gradient of φ is the direction of steepest ascent for maximize φ. By multiplying the gradient by t inverse of the Hessian matrix of φ we have the search direction of the Newton-Raphson technique. F k large enough

$$\lambda_i^{k+1} = \lambda_i^k [\exp [r_k g_i(x_k) - 1]] + \lambda_i^k. \tag{31}$$

Since $\Lambda^{-1}(\lambda^k)$ approximates the Hessian of $\varphi(\lambda^K)$ for k large enough, $\Lambda(\lambda^k)$ approximates the verse of the Hessian. This means that the direction of improvement in going from λ_i^k to λ_i^{k+1} in (31) approximately that of the Newton-Raphson technique. This property is termed *asymptotic Newton* [8].

CONCLUSIONS

The importance of an algorithm having the asymptotic Newton structure comes about in t following manner. The rate of convergence of steepest ascent is linear, and the rate improvement a function of the ratio of the difference of the largest and smallest eigenvalues of the Hessian to t sum of these eigenvalues. In the algorithm presented here we have steepest ascent search directio on $\varphi(\lambda)$ within a vector space translated by an estimate of the size of the multipliers. Within t transformed space, the Hessian more closely approximates the identity matrix as r_k is increase

ıging the largest and smallest eigenvalues closer together. This means that the factor that governs rate of convergence is improving as r_k increases. As a consequence an asymptotic Newton algorithm a superlinear convergence rate.

Whether this algorithm is superior to the algorithms that do not have the second order differentia-y property is not clear. There is more computational effort at each iteration because nonbinding straints still can affect the calculation of the Hessian for the unconstrained minimization long after would have been dropped with the other algorithms. Also, r_k must be increased to infinity in the t. The choice of algorithm should be dependent on whether the discontinuity of the Hessian is a blem.

BIBLIOGRAPHY

] Bertsekas, Dimitri, "On Penalty and Multiplier Methods for Constrained Minimization," Department of Electrical Engineering, working paper, University of Illinois at Urbana, Champaign (Apr. 1974).

] Bertsekas, Dimitri, "Combined Primal-Dual and Penalty Function Methods for Constrained Minimization," SIAM J. Control *13* (1975).

Evans, J. P. and F. J. Gould, "An Existence Theorem for Penalty Function Theory," SIAM J. Control *12* (1974).

Fiacco, A. V. and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques* (Wiley, New York, 1968).

Hestines, M. R., "Multiplier and Gradient Methods," J. Optimization Theory and Appl. *4*, 303–320 (1969).

Luenberger, D. G., *Introduction to Linear and Nonlinear Programming* (Addison-Wesley, Reading, Mass., 1973).

Mangasarian, O. L., *Nonlinear Programming* (McGraw-Hill, New York, 1969).

Mangasarian, O. L., "Unconstrained Methods in Optimization," Computer Sciences Technical Report #224, University of Wisconsin (1974).

Murphy, F. H., "A Class of Exponential Penalty Functions," SIAM J. Control *12* (1974).

Murphy, F. H., "Topics in Nonlinear Programming, Penalty Function and Column Generation Algorithms," Ph. D. Dissertation, Yale University (1971).

Murphy, F. H., "A Generalized Lagrange Multiplier Function Algorithm for Nonlinear Programming," Discussion paper 114, Center for Mathematical Studies in Economics and Management Science (Oct. 1974).

O'Neill, R. P. and W. B. Widhelm, "Acceleration of Lagrangian Coleman-Generation Algorithms by Penalty Function Methods," University of Maryland—College Park (Jan. 1975).

Powell, M. J. D., "A Method of Nonlinear Constraints in Minimization Problems," in *Optimization*, R. Fletcher, ed. (Academic Press, New York, 1969).

ESTIMATION OF STRATEGIES IN A MARKOV GAME

Jerzy A. Filar
Monash University
Melbourne, Australia

ABSTRACT

In this paper a two-person Markov game, in discrete time, and with perfect state information, is considered from the point of view of a single player (player A) only. It is assumed that A's opponent (player B) uses the same strategy every time the game is played. It is shown that A can obtain a consistent estimate of B's strategy on the basis of his past experience of playing the game with B. Two methods of deriving such an estimate are given. Further, it is shown that using one of these estimates A can construct a strategy for himself which is asymptotically optimal. A simple example of a game in which the above method may be useful is given.

ATURE OF THE PROBLEM

Ve shall consider a Markov game (players A and B) with perfect state information and in discrete from the point of view of player A only. A's objective is to maximize his expected payoff func- (x, y), where x denotes A's own strategy and y denotes B's strategy. We impose the following tions:

) The players have no knowledge of their opponent's strategy or payoff functions.

i) A knows that B uses the same strategy, η, every time the game is played.

learly, it would be in A's interest to gather information about B's strategy, which would help him choice of his own strategy. Hence the problem is to devise an estimation procedure which will player A to form a reliable estimate, $\hat{\eta}$, of B's true strategy η. Now, by maximizing the function) with respect to x, A could obtain a strategy x^* (not necessarily unique) for himself, which we equire to be asymptotically optimal subject to B continuing to use η. That is, if $\hat{\eta}$ is based on a number of past plays, then $v(x^*, \eta)$ will approximate better to $\max_{(x)} v(x, \eta)$. The maximization all possible strategies for A.

ELATIONSHIP TO EXISTING THEORIES

1967 J. C. Harsanyi [2] developed a new approach to games with incomplete information. He d that under certain conditions such games can be reduced to equivalent games with complete ation involving some random elements, for which Nash equilibrium points can be found. Har- theory initiated research into repeated games of incomplete information by R. J. Aumann, schler, J. F. Mertens, S. Zamir and others; for example, see [1], [3], [7], [8], and [9]. These are in which each player has only partial information about a chance move that takes place at the ing of the game. Two alternative approaches were applied to the often-repeated games of this he first is to consider the n-times repeated game Γ_n and its value v_n and then find out $\lim_{n\to\infty} v_n$ if s. The second approach is to treat directly the infinitely repeated game Γ_∞ and find its value

v_∞ when it exists. The rate of convergence of v_n to its limit has also been discussed in some c

Further, in 1969 J. E. Walsh in his "Median Game Theory" (see [5]) recognized the fact that librium point solutions may not be satisfactory criteria for optimality in many game situations su economic and military games. He differentiated between situations where the players behave tectively and/or vindictively, since assumptions about the players' motivation affect the choice optimality criterion.

The game considered in this paper is a multistage game with incomplete information in wh special assumption is made about the behavior of one player, and thus it resembles the games tioned above. However, no attempt is made here to solve the game in the equilibrium strategy s Instead, a statistical technique is proposed for estimating the "missing information" as well as a nique for utilizing estimates thus obtained to construct a strategy for one player which is better the minimax strategy. Further, we show that the strategy based on our estimates converges in p bility to the "optimal" strategy for one player, provided that a special assumption about the pl holds. This approach is motivated by the idea that when playing a game against a particular opp one can benefit from prior information about that player's strategy. The method suggested in paper is straightforward and effective in practical applications.

3. A MARKOV GAME IN DISCRETE TIME

We shall use Zachrisson's definition of a Markov game (see [6]) with some modifications. A sl of its structure is given below.

Suppose we have a two-person game which at any time of play is in one and one only of N ic fiable states, we number them $1, 2, \ldots, N-1, N$. We shall assume a finite time of play, T^*, and in reverse time, i.e., the game starts at T^* and ends at 0. Players A and B choose their strategy pa eters at discrete points of time $T^*, T^*-1, \ldots, t+1, t, \ldots 2, 1$. Thus, if the game is in state i at $t+1$, A and B will choose a pair of strategy parameters $x_i(t+1)$, $y_i(t+1)$ respectively. We note the players' strategies are functions of both time and state. The game is then transferred to an state where it will remain until the next decision time at the reverse time t. We write,

$$p_{ij}(t)=p_{ij}(t;\, x_i(t+1),\, y_i(t+1)) \tag{3.1}$$

$$=Pr\{\text{game goes to state } j \text{ at time } t/\text{it was in state } i \text{ at } t+1 \text{ and } A \text{ and } B \text{ used strat}$$
$$x_i(t+1),\, y_i(t+1) \text{ resp.}\}$$

where $j \epsilon S=\{1, 2, \ldots, N\}$. The corresponding transition matrix can be written as:

$$\underset{\approx}{P}(t)=\underset{\approx}{P}(t;\, \underset{\sim}{x}(t+1),\, \underset{\sim}{y}(t+1))=(p_{ij}(t))_{i,j=1}^{N} \tag{3.2}$$

where the vector

$$\underset{\sim}{x}(t+1)=\begin{pmatrix} x_1(t+1) \\ x_2(t+1) \\ \cdot \\ \cdot \\ \cdot \\ x_N(t+1) \end{pmatrix}$$

es A's complete strategy at time $t+1$. Vector $\underset{\sim}{y}(t+1)$ is defined similarly. Thus the game is

nined by a sequence of T^* matrices: $\underset{\approx}{P}(T^*-1), \ldots, \underset{\approx}{P}(t), \ldots, \underset{\approx}{P}(2), \underset{\approx}{P}(1), \underset{\approx}{P}(1), \underset{\approx}{P}(0)$.

/e deviate from Zachrisson in the definition of the players' payoff functions. We define:

$_{ij}^{(A)}(t) = w_{ij}^A(t; x_j(t+1), y_i(t+1))$ as A's payoff in state j at time t, provided that game was in state i at time $t+1$ and players' strategies were $x_i(t+1)$ and $y_i(t+1)$ at that time. Where $t=1, \ldots, T^*-1$.

he matrix:

$$\underset{\approx}{W}^A(t) = (w_{ij}^A(t))_{i,j=1}^N$$

ed player A's *current payoff matrix* at t. We define a similar payoff matrix $\underset{\approx}{W}^B(t)$ for player B, general we shall assume that $w_{ij}^A(t) + w_{ij}^B(t) \neq 0$. Further, we let:

(t) = player A's expected payoff from time t until the end of the game, if the game is in state i at time t, where $t=1, 2, \ldots, T^*$ and we call the vector $\underset{\sim}{v}^A(t) = (v_1^A(t), v_2^A(t), \ldots, v_N^A(t))^T$ player A's *value vector at time t*, with $\underset{\sim}{v}^B(t)$ defined similarly.

e would expect the value vectors at successive times to be related to one another by some kind ırsive equation. It turns out that if we define the matrix operator $\mathscr{D}$ by:

$$\mathscr{D}(A) = (a_{11}a_{22}, \ldots, a_{NN})^T = \text{a vector of diag. elements of } A \text{ where } A \text{ is any matrix } A = (a_{ij})_{i,j=1}^N$$

e following backward equations hold:

$$\underset{\sim}{v}^A(1) = \mathscr{D}\{[\underset{\approx}{P}(0)][\underset{\approx}{W}^A(0)]^T\}$$

$$\underset{\sim}{v}^A(t+1) = \mathscr{D}\{[\underset{\approx}{P}(t)][\underset{\approx}{W}^A(t)]^T\} + \underset{\approx}{P}(t)\underset{\sim}{v}^A(t)$$

$= 1,2, \ldots, T^*-1$. Equations (3.5) simply state that for each i,

$$v_i^A(t+1) = \sum_{j=1}^N p_{ij}(t)[w_{ij}^A(t) + v_j^A(t)]$$

s what we would expect. We shall define the distribution vector at time t by: $\underset{\sim}{q}(t) = (q_1(t), q_2$ $., q_1(t))^T$ for $t=0, 1, \ldots, T^*-1, T^*$, where $q_i(t) = Pr\{\text{game is in state } i \text{ at time } t\}$.
umed that the initial distribution vector $\underset{\sim}{q}(T^*)$ is known. It is easily verifiable that:

$$\underset{\sim}{q}(t) = [\underset{\approx}{P}(t)]^T\underset{\sim}{q}(t+1) \qquad \text{for } t=T^*-1, \ldots, 2, 1, 0.$$

Since in the remainder of this paper we shall consider only player A's payoffs, the superscript A in $v_i^A(t)$, $w_{ij}^A(t)$, $\underset{\approx}{W}^A(t)$ and $\underset{\sim}{v}^A(t)$ will be omitted from now on.

4. DERIVATION OF ESTIMATES OF B's STRATEGY

Postulate (ii) of section 1 ensures that every time the game is in state i at the $(t+1)$th decision point, player B uses a strategy parameter η_i. Suppose that in order to estimate η_i A plays the game with B n times and keeps a record of all transitions which eventuate. We shall assume that during those n plays A also uses a fixed strategy, ξ. ξ consists of a set of T^* vectors $\underset{\sim}{\xi}(t)$.

Let K_i = the number of times the game was in state i at $(t+1)$th decision point,

and S_{ij} = the number of times the transition $i \to j$ was observed in time $t+1 \to t$.

Clearly

$$\sum_{i=1}^{N} K_i = n \text{ and } \sum_{j=1}^{N} S_{ij} = K_i.$$

K_i and S_{ij} are random variables with the distributions $B_i(n, q_i(t))$ and $B_i(K_i, p_{ij}(t))$ respectively. Now, since A knows his own strategy, $p_{ij}(t)$, as in (3.1), is a function of η_i only. Let us write $p_{ij}(t) = f_j(\eta_i)$ where η_i will usually be restricted to some interval $[a_i, b_i]$ by the nature of the game. It is natural for A to take the statistic

$$\frac{S_{ij}}{K_i} \text{ to be an estimator of } p_{ij}(t).$$

Then an estimate $\hat{\eta}_i$ of η_i can be found by solving the equation:

$$\frac{S_{ij}}{K_i} = f_j(\hat{\eta}_i) \qquad \text{for } \hat{\eta}_i. \tag{4.1}$$

Further, if f_j is a one-to-one function we can write

$$\hat{\eta}_i = g_j\left(\frac{S_{ij}}{K_i}\right) \qquad \text{where } g_j = f_j^{-1}, \tag{4.2}$$

the inverse function of f_j. It should be clear that with this approach any one of the functions f_j ($j=1, 2, \ldots, N$) could be used for the purpose of estimating η_i. In some cases the average of the estimates thus obtained could be appropriate.

An alternative method of deriving an estimate of η_i is by the method of maximum likelihood. If the vector of observations for a given K_i be: $\underset{\sim}{S} = (S_{i1}, S_{i2}, \ldots S_{iN}) \in \mathcal{W} \subset \mathcal{R}^N$ where it is possible to list all elements of $\mathcal{W}$. Let us write $y_i = y_i(t+1) \in [a_i, b_i]$ where η_i is the true value of y_i. We shall define $\hat{\hat{\eta}}_i$ to be the *m.l.e.* of η_i if:

$$L(\hat{\hat{\eta}}_i | K_i, \underset{\sim}{S}) = \sup_{y_i \in [a_i, b_i]} L(y_i | K_i, \underset{\sim}{S}), \tag{4.3}$$

ere

$$L(y_i|K_i, \underset{\sim}{S}) = \frac{K_i!}{S_{i1}!S_{i2}! \ldots S_{iN}!} \prod_{j=1}^{N} [p_{ij}(y_i)]^{S_{ij}}$$

he likelihood function since the joint distribution of the S_{ij}'s is clearly multinomial. The following ılts are special cases of the results proved in Rao [4], on pages 293–298.

a. If, (1) the functions $p_{ij}(y_i)$ for $j=1, 2, \ldots, N$ are continuous in y_i on $[a_i, b_i]$, and (2) Given >0, there exists ϵ:

$$\inf_{|y_i-\eta_i|>\delta} \left\{ \sum_{j=1}^{N} p_{ij}(\eta_i) \log \left[\frac{p_{ij}(\eta_i)}{p_{ij}(y_i)} \right] \right\} \geq \epsilon$$

ı, a *m.l.e.* $\hat{\hat{\eta}}_i$ exists and converges to η_i with probability 1 as $K_i \to \infty$.

b. If the condition (2) of (a) holds and the functions $p_{ij}(y_i)$ admit 1st order derivatives which are tinuous, then the *m.l.e.* $\hat{\hat{\eta}}_i$ is asymptotically normally distributed.

Should we desire to construct a confidence interval for η_i we could use the normal approximation he distribution of $\hat{\hat{\eta}}$. Otherwise, for any $y_i \epsilon [a_i, b_i]$ we could construct a region $W_\alpha(y_i) \subset W$ such :

$$Pr\{\underset{\sim}{S} \epsilon W_\alpha(y_i)|y_i\} > 1-\alpha.$$

Then if we observe a vector $\underset{\sim}{S}^* \epsilon W$ such that $\underset{\sim}{S}^* \epsilon W_\alpha(y_i)$ for $y_i \epsilon [l_1, l_2] \subseteq [a_i, b_i]$, we can state :

$$Pr\{l_1 \leq \eta_i \leq l_2 | \underset{\sim}{S}^*\} > 1-\alpha.$$

It is also worth mentioning that estimators S_{ij}/K_i and $\hat{\eta}_i$ derived in (4.1) and (4.2) are consistent $_{ij}(t)$ and η_i respectively. The distribution of these estimators is difficult to determine; however, moments can be either found exactly or be approximated to by standard techniques. In particular define $E\left[\frac{S_{ij}}{K_i} \,\middle|\, K_i=0\right] = p_{ij}(t)$, then

$$\text{Var}\left(\frac{S_{ij}}{K_i}\right) \approx \frac{1}{n} [p_{ij}(t)]^2 \left(\frac{1}{q_i(t+1)p_{ij}(t)} - \frac{1}{q_i(t+1)} \right)$$

this $\to 0$ as $n \to \infty$. Further, if we assume that $g_j = f_j^{-1}$ and its first two derivatives $g_j^{(1)}$ and $g_j^{(2)}$ and are bounded at the point $p_{ij}(t) = E\left[\frac{S_{ij}}{K_i}\right]$, then it follows that:

$$E(\hat{\eta}_i) = E\left[g_j\left(\frac{S_{ij}}{K_i}\right)\right] \text{ by (4.2)}$$

$$\approx \eta_i + \frac{1}{2} \cdot g_j^{(2)}(p_{ij}(t)) \cdot \text{Var}\left(\frac{S_{ij}}{K_i}\right),$$

and

$$\text{Var}\,(\hat{\eta}_i) \approx \text{Var}\left(\frac{S_{ij}}{K_i}\right)\left\{[g_j^{(1)}(p_{ij}(t))]^2 - \frac{1}{4}\,[g_j^{(2)}(p_{ij}(t))]^2\,\text{Var}\left(\frac{S_{ij}}{K_i}\right)\right\}. \tag{4.9}$$

Thus $E(\hat{\eta}_i) \to \eta_i$ and Var $(\hat{\eta}_i) \to 0$ as $n \to \infty$ and $\hat{\eta}_i$ is a consistent but biased estimator of η_i.

A confidence interval for $p_{ij}(t)$ can be obtained using the fact that for $K_i = k_i$ the random variat

$$Z = \frac{S_{ij} - k_i p_{ij}(t)}{\sqrt{k_i p_{ij}(t)\,(1 - p_{ij}(t))}} \text{ approx. } N(0, 1). \tag{4.10}$$

So, from normal distribution we can find t_α: $Pr\{|Z| < t_\alpha\} = 1 - \alpha$, which can be written as:

$$Pr\{l_1(S_{ij}|k_i) < p_{ij}(t) < l_2(S_{ij}|k_i)\} = 1 - \alpha \tag{4.11}$$

where l_1, l_2 equal

$$\frac{(2S_{ij} + t^2) \mp \sqrt{t_\alpha + 4S_{ij}t_\alpha - \dfrac{4S_{ij}t_\alpha{}^2}{k_i}}}{2(t_\alpha{}^2 + k_i)} \qquad \text{respectively.}$$

Further, if the function $g_j = f_j{}^{-1}$ is monotone increasing, we have that:

$$Pr\{g_j[l_1(S_{ij}|k_i)] < \eta_i < g_j[l_2(S_{ij}|k_i)]\} = 1 - \alpha, \tag{4.12}$$

with the inequalities above simply reversed if g_j were monotone decreasing. Note that the width intervals in (4.11) and (4.12) tends to zero as $k_i \to \infty$, provided that g_j is continuous as well.

COMMENTS:

1. Confidence limits for η_i conditional on $K_i = k_i$ are even more useful than if they were con tional on n, for to estimate B's strategy in ith state at the $(t+1)$th stage, player A would want to o serve a sufficiently large value for K_i.

2. Note that since we were considering a typical time interval $(t+1) \to t$, the argument (t) w dropped in this section. Strictly, we should have written; $\eta_i(t+1)$, $K_i(t+1)$, $S_{ij}(t)$, etc.

3. To obtain an estimate of B's complete strategy, the estimation procedure outlined above wou have to be repeated for each of the states 1, 2, . . ., N and at each decision point T^*, T^*-1, . . . t, . . ., 2, 1.

5. DERIVATION OF A's STRATEGY

Let us represent B's true strategy for the whole game, η, by a set of T^* vectors:

$$\underset{\sim}{\eta}(t) = \begin{pmatrix} \eta_1(t) \\ \eta_2(t) \\ \vdots \\ \eta_N(t) \end{pmatrix}, \qquad t = 1, 2, \ldots, T^*.$$

We shall define player A's *optimal strategy*, x^0, as the set of vectors $\underset{\sim}{x}^0(t); t = 1, 2, \ldots, T^*$ which sfies the following equations:
) see (3.5))

$$\underset{\sim}{v}^0(1) = \underset{\sim}{v}(\underset{\sim}{x}^0(1), \underset{\sim}{\eta}(1))$$

$$= \max_{(x(1))} \mathscr{D}\{\underset{\approx}{P}(0; \underset{\sim}{x}(1), \underset{\approx}{\eta}(1))[\underset{\approx}{W}(0; \underset{\sim}{x}(1))]^T\}$$

$$+1) = \underset{\sim}{v}(\underset{\sim}{x}^0(t+1), \underset{\sim}{\eta}(t+1))$$

$$= \max_{(x(t+1))} \{\mathscr{D}\{[\underset{\approx}{P}(t; \underset{\sim}{x}(t+1), \underset{\sim}{\eta}(t+1)[\underset{\approx}{W}(t)]^T] + \underset{\approx}{P}(t; \underset{\sim}{x}(t+1), \underset{\sim}{\eta}(t+1))\underset{\sim}{v}^0(t)\}\}$$

$= 0, 1, 2, \ldots, T^* - 1$, the maximization over $(\underset{\sim}{x}(t+1))$ means that the ith row equation in (5.2) is imized over $x_i(t+1)$ which will usually be restricted to a bounded interval $[c_i, d_i]$ specified by the s of the game.

Suppose now, that by following an estimation procedure of section 4, A obtains an estimate, $\hat{\eta}$, sisting of vectors $\hat{\underset{\sim}{\eta}}(t) t = 1, \ldots, T^*$) of B's strategy. Then by following the maximization process 5.2) with vectors $\underset{\sim}{\eta}(t)$ replaced by $\hat{\underset{\sim}{\eta}}(t)$, A will obtain a strategy, x^*, (consisting of vectors $\underset{\sim}{x}^*(t)$, 2, . . ., T^*) and a new set of expected payoff vectors: $\underset{\sim}{v}^*(t) = \underset{\sim}{v}(\underset{\sim}{x}^*(t), \underset{\sim}{\eta}(t)) t = 1, 2, \ldots, T^*$. ıe next section a comparison will be made between the effects of employing the strategies x^* and x^0.

Further, if initially A has no idea as to what η really is, he might adopt a pessimistic approach, assume that B plays directly against him. Thus in Equations (3.5) for each admissible $\underset{\sim}{y}(t)$ he can : out

$$\underset{\sim}{V}(y(t)) = \max_{\underset{\sim}{x}(t)} \underset{\sim}{v}(\underset{\sim}{x}(t), \underset{\sim}{y}(t)) \qquad t = 1, \ldots, T^*$$

h yields a vector, ${}_1\underset{\sim}{x}(t) = {}_1\underset{\sim}{x}(\underset{\sim}{y}(t))$ such that

$$\underset{\sim}{V}(y(t)) = \underset{\sim}{v}({}_1\underset{\sim}{x}(t), \underset{\sim}{y}(t)) \qquad t = 1, \ldots, T^*.$$

Then according to the pessimistic assumption A thinks that B is using a strategy, γ, consisting of ve tors $\underset{\sim}{\gamma}(t)$ $t=1, \ldots, T^*$ such that:

$$\underset{\sim}{V}(\underset{\sim}{\gamma}(t)) = \min_{\underset{\sim}{\gamma}(t)} \underset{\sim}{V}(\underset{\sim}{\gamma}(t)) \qquad t=1, \ldots, T^* \tag{5.5}$$

Thus, during the experimental plays of the game A will employ a strategy, ξ, consisting of vect $\underset{\sim}{\xi}(t) = {}_1\underset{\sim}{x}(\underset{\sim}{\gamma}(t))$ $t=1, \ldots, T^*$ and he will expect his value vectors to be:

$$\begin{aligned} \underset{\sim}{V}(\underset{\sim}{\gamma}(t)) &= \underset{\sim}{v}(\underset{\sim}{\xi}(t), \underset{\sim}{\gamma}(t)) \\ &= \min_{\underset{\sim}{\gamma}(t)} \max_{\underset{\sim}{x}(t)} \underset{\sim}{v}(\underset{\sim}{x}(t), \underset{\sim}{\gamma}(t)) \end{aligned} \tag{5.6}$$

for $t=1, 2, \ldots, T^*$. Whereas in fact his value vectors will be given by

$${}_1\underset{\sim}{v}^a(t) = \underset{\sim}{v}(\underset{\sim}{\xi}(t), \underset{\sim}{\eta}(t)) \qquad t=1, \ldots, T^* \tag{5.7}$$

Similarly, if A employs x^*, his actual value vectors will not be $\underset{\sim}{v}^*(t)$'s but:

$$\underset{\sim}{v}^a(t) = \underset{\sim}{v}(\underset{\sim}{x}^*(t), \underset{\sim}{\eta}(t)) \qquad t=1, \ldots, T^* \tag{5.8}$$

6. EVALUATION OF THE ESTIMATION TECHNIQUE

Clearly, the strategy, x^*, based on the estimate $\hat{\eta}$ will be *asymptotically optimal* if the vect

$$*\underset{\sim}{v}^a(t) \xrightarrow{pr} \underset{\sim}{v}^0(t) \qquad \text{for} \qquad t=1, 2, \ldots, T^*$$

We shall show that this is in fact so under quite general conditions. Consider the ith row of Equatio (5.2)–(5.8) only. If $\hat{\eta}$ is obtained as in section 4, we have that

$$\hat{\eta}_i(t) \xrightarrow{pr} \eta_i(t) \qquad \text{as } k_i(t) \to \infty.$$

Now, if we assume that $v_i(x_i(t), y_i(t))$ is jointly continuous in $x_i(t)$ and $y_i(t)$, then it can be sho that $V_i(y_i(t))$ is continuous as well and hence that

$$v_i^*(t) = V_i(\hat{\eta}_i(t)) \xrightarrow{pr} V_i(\eta_i(t)) \qquad \text{as } k_i(t) \to \infty, \tag{6.1}$$

but

$$v_i^a(t) - v_i^*(t) = v_i(x_i^*(t), \eta_i(t)) - v_i^*(x_i^*(t), \hat{\eta}_i(t)) \xrightarrow{pr} 0 \qquad \text{as } k_i(t) \to \infty; \tag{6.2}$$

therefore,

$$v_i^a(t) \xrightarrow{pr} V_i(\eta_i(t)) = v_i^0(t) \qquad \text{as } k_i(t) \to \infty$$

equired. Further, by looking at Equations (3.5), we see that $v_i(x_i(t), y_i(t))$ will be continuous if functions $p_{ij}(t-1, x_i(t), y_i(t))$ and $W_{ij}(t-1, x_i(t), y_1(t))$ are continuous. Note that all state-ts about continuity need to refer only to the intervals of interest specified by the rules of the game. e the recursive equations involve only finitely many steps, the statements about asymptotic opti-y can be proved by induction for each $t=1, \ldots, T^*$.

In practice, when playing the experimental games, player A will want to know "how soon" can he using, x^*, with a reasonable confidence that this will prove better than, ξ. For a typical i and t can be illustrated as follows in Figure 1:

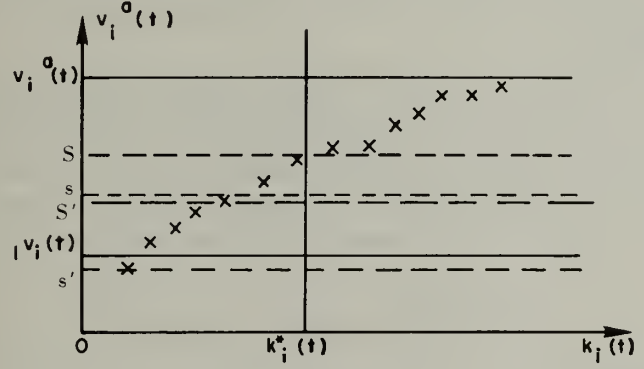


FIGURE 1.

Thus, player A requires a criterion for deciding how large a value of $k_i(t)$ needs be observed before n assert with at least $(1-\alpha)100$ percent confidence that the use of $x_i^*(t)$ is better than of $\xi_i(t)$, $v_i^a(t) > {}_1v_i^a(t)$). One such criterion is derived below.

Suppose that when estimating $\eta_i(t)$ we construct a $(1-\alpha)100$ percent C.I. $[L_1, L_2] = I$ as in . Let us define the following quantities

$$S = \max_{y_i(t) \in I} v_i(x_i^*(t), y_i(t)) \qquad \text{and } s = \min_{y_i(t) \in I} v_i(x_i^*(t), y_i(t))$$

$$S' = \max_{y_i(t) \in I} v_i(\xi_i(t), y_i(t)) \qquad \text{and } s' = \min_{y_i(t) \in I} v_i(\xi_i(t), y_i(t)),$$

$$Pr\{s \leqslant^* v_i^a(t) \leqslant S\} > 1-\alpha$$

$$Pr\{s' \leqslant {}_1v_i^a(t) \leqslant S'\} > 1-\alpha.$$

since the width of $I \to 0$ as $k_i(t) \to \infty$ we can find $k_i^*(t)$ for which $S' < s$ (see fig. 1), then with bility of at least $1-\alpha$ the use of the estimation strategy $x_i^*(t)$ will yield a better expected payoff

than the minimax strategy $\xi_i(t)$. Hence a complete set $K=\{k_i^*(t)\colon i=1, \ldots, N \text{ and } t=1, \ldots,$ can be constructed and a corresponding general stopping could be: A continues to use, ξ until ev member of K has been observed.

COMMENTS:

1. If the numbers T^* and N are large, quite clearly a very large number of observations may required. However, if some states are very unlikely, player A may not bother to estimate B's strat in those states.

2. A major drawback of the above approach is that player B might change his strategy soon a A starts using x^*. Thus the benefit to A could be short-lived. On the other hand, if B is a "conservati player or if his returns remain satisfactory this may not happen.

7. A SIMPLE EXAMPLE †

Suppose we have two contractors, A and B, bidding at discrete points of time for a contract (wh has to be renewed monthly, for instance). Let state 1 correspond to B getting the contract for a mor and let state 2 correspond to A getting it. Players' strategies in state i are quotes (scaled down) for contract x_i, $y_i \in [0, 1]$ $i=1, 2,$. We consider a single-decision game, i.e. $T^*=1$. Suppose that

$$\underset{\approx}{P}(0)=\frac{1}{2}\begin{pmatrix}1+x_1-y_1 & 1-x_1+y_1\\ 1+x_2-y_2 & 1-x_2+y_2\end{pmatrix}, \tag{7.1}$$

and that A's payoff matrix is:

$$\underset{\approx}{W}^A(0)=\begin{pmatrix}0 & cx_1\\ 0 & cx_2\end{pmatrix} \quad \text{where } c>1, \tag{7.2}$$

i.e. if A doesn't get the contract, his payoff is 0, and if he gets it his payoff is c times the quote he ga (which was scaled down). Then from Equation (3.5) we have that A's expected payoff vector is:

$$\underset{\sim}{v}(1)=\begin{pmatrix}v_1(x_1, y_1)\\ v_2(x_2, y_2)\end{pmatrix}=\frac{1}{2}\,c\,\begin{pmatrix}x_1+x_1y_1-x_1^2\\ x_2+x_2y_2-x_2^2\end{pmatrix}. \tag{7.3}$$

Let player B's strategy (unknown to A) be represented by the vector $\underset{\sim}{\eta}=\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix}$. We shall assume th the game starts in state 1. Thus A is interested in estimating η_1 only, and in maximizing $v_1(x_1, \eta$ Now, the maximum of $\frac{1}{2}c(x_1+x_1y_1-x_1^2)$ for a given y_1 occurs when $x_1=\frac{1}{2}(1+y_1)$, so we have fr (5.3) that

† This example is a special case of a general Markov game model of duopoly which I have constructed recently. There a number of reasonable ways of analyzing the general model, depending largely on the assumptions about players' motiva What is done here, however, is intended purely as an illustration of our statistical technique.

$$V_1(y_1) = \max_{x_1 \in [0,1]} v_1(x_1, y_1) = \frac{1}{8} c(1+y_1)^2$$

$$V_1(0) = \min_{(y_1)} V_1(y_1) = \frac{1}{8} c, \text{ and that}$$

0 while $\xi_1 = {}_1x_1(0) = \frac{1}{2}(1+0) = \frac{1}{2}$. Thus during the experimental plays A will use $\xi_1 = \frac{1}{2}$. While he
ng $\xi_1 = \frac{1}{2}$, the function $f_1(y_1) = p_{11}(\xi_1, y_1) = \frac{1}{2}\left(\frac{3}{2} - y_1\right)$, so the function $g_1(z) = f_1^{-1}(z) = \frac{3}{2} - 2z$
tinuous and monotone decreasing. Thus if we use f_1 to estimate η_1 we have as in (4.2) that:

$$\hat{\eta}_1 = \frac{3}{2} - 2\frac{S_{11}}{K_1} .$$

that in this case S_{11} determines S_{12} as well, for a given K_1, so that the *m.l.e.* $\hat{\hat{\eta}}$ will not be any better
$\hat{\eta}_1$ above.)
Now, suppose that after 40 plays (each starting in state 1), A observed $s_{11} = 10$, $s_{12} = 30$ ($K_1 = 40$
rse), then $\hat{\eta}_1 = \frac{3}{2} - 2\frac{1}{4} = 1$, and if he requires a 90% C.I. for $\hat{\eta}_1$, he can take $\alpha = .1$, and by the
d of section 4 he will find that $Pr\{.1709 \leq p_{11} \leq .3608\} = .9$ and similarly that, $Pr\{.7784 \leq \eta_1 \leq 1\}$

rom here A obtains x_1^* by finding

$$\max_{x_1 \in [0,1]} v_1(x_1, \hat{\eta}_1) = \max_{(x_1)} \left\{\frac{1}{2} c(2x_1 - x_1^2)\right\} = \frac{c}{2}$$

$x_1 = x_1^* = 1$, and his actual expected payoff when he uses $x_1^* = 1$ will be: ${}_*v_1{}^a = v_1(x_1^*, \eta_1) = \frac{1}{2} c \cdot \eta_1$.
hus by the 1st equation of (6.4) $Pr\{.3892c \leq {}_*v_1^a \leq .5000c\} > .9$. Whereas if A uses the minimax
y $\xi_1 = \frac{1}{2}$ his expected payoff will actually be: ${}_1v_1^a = v_1(\xi_1, \eta_1) = c\left(\frac{1}{8} + \frac{\eta_1}{4}\right)$.
, by the 2nd equation of (6.4) $Pr\{.3196c \leq {}_1v_1^a \leq .3750c\} > .9$, which satisfies the stopping rule
ie number of trials as small as 40.

owledgment

would like to thank Professor J. S. Maritz of Monash University with whom I discussed the
its of this paper on a number of occasions.

REFERENCES

[1] Aumann, R. J. and M. Maschler, "Repeated Games with Incomplete Information. A survey of cent results". Report to the U.S. Arms Control and Disarmament Agency, Washington, D Final report on contract ACDA/ST–143, Ch. III (Mathematica, Princeton, N.J., Sept. 1967).

[2] Harsanyi, J. C., "Games with Incomplete Information Played by 'Bayesian' Players: Parts I, III." Man. Sci., **14** (1967).

[3] Mertens, J. F. and S. Zamir, "The Value of Two-Person Zero-Sum Repeated Games" Internatio J. Game Theory **1** (1971).

[4] Rao, C. R., *Linear Statistical Inference and its Applications* (Wiley and Sons, New York, 19

[5] Walsh, J. E., "Discrete Two-Person Game Theory with Median Payoff Criterion" Opsearch, **6** (19

[6] Zachrisson, L. E., " 'Markov Games' Advances in Game Theory" Annals of Maths. Studies. St 52 (Princeton University Press, 1964).

[7] Zamir, S., "On the Relation between Finitely and Infinitely Repeated Games with Incomplete formation" International J. Game Theory **1** (1971).

[8] Zamir, S., "On Repeated Games with General Information Function". International J. Game The **2** (1973).

[9] Zamir, S., "On the Notion of Value for Games with Infinitely Many Stages". The Annals of S tistics **1** (1973).

LOWSHOP SEQUENCING PROBLEM WITH ORDERED PROCESSING TIME MATRICES: A GENERAL CASE*

M. L. Smith, S. S. Panwalkar, and R. A. Dudek

Texas Tech University
Department of Industrial Engineering
Lubbock, Texas

ABSTRACT

The ordered matrix flow shop problem with no passing of jobs is considered. In an earlier paper, the authors have considered a special case of the problem and have proposed a simple and efficient algorithm that finds a sequence with minimum makespan for a special problem. This paper considers a more general case. This technique is shown to be considerably more efficient than are existing methods for the conventional flow shop problems.

RODUCTION

In [4], Smith, Panwalkar and Dudek have defined a subcategory of the n job, m machine problem d an "ordered flowshop problem." A flowshop problem is called an ordered problem if the following roperties are satisfied:

1) If a particular job has a smaller processing time on any machine than does a second job on the machine, this implies then that the processing time of this first job is less than or equal to the ssing time of the second job on all corresponding machines.

2) If one job has its jth smallest processing time on some machine, this implies that every other ill have its jth smallest processing time on the same machine. Beyond these special properties e processing times, the usual assumptions for the general flowshop problems [1] also apply to the ed matrix problem. One important assumption is that of no passing of jobs.

In an earlier paper Smith, Dudek and Panwalkar [3] first introduced the ordered flowshop problem. enumerated several hundred problems and observed two interesting characteristics for the ed problem. First, if the maximum processing time for every job in an ordered problem occurred e first (last) machine, then the sequence obtained by arranging jobs in descending (ascending) of processing times always represented a minimum makespan sequence.

Secondly, if the maximum processing time for every job in an ordered flowshop problem occurred intermediate machine, then there was always a minimum makespan sequence such that a subset os was arranged in ascending order of processing times followed by the remaining jobs in de- ling order.

Based on these observations Smith, et al. [3] proposed two algorithms for the ordered flowshop em, one for the special case when the maximum processing time for every job occured on the first

his research was partially supported by the National Science Foundation Grant GK 2869.

or the last machine, and the second for the general case. Although complete enumeration showed a all problems attempted were optimally solved by the algorithms, no proofs of optimality were develo

Recently, Smith et al. [4] were able to develop the proof of optimality for the algorithm for e special case of the ordered flowshop problem. In this paper we will present the second algorithm d show that the algorithm will generate at least one optimal solution.

PROPOSED ALGORITHM

To solve the ordered matrix problem the following special algorithm was developed. This meth uses a form of limited enumeration with 2^{n-1} sequences being formed. The steps for the algorithm a

Step 1. Rank the jobs according to ascending processing times on the first machine. (In case of ties, break the ties by considering the processing times on the second or third machine if necessa

Step 2. Place the lowest ranking job which has not been sequenced into the leftmost unfilled quence position in each partial sequence being constructed. Also place this same job in the rightmo unfilled sequence position in each partial sequence. Note that the number of partial sequences doub as each job (other than the highest ranking job) is added.

Step 3. Repeat Step 2 until the first $n-1$ jobs are placed into every sequence; then the high ranking job is to be placed into the only unfilled position.

Step 4. The 2^{n-1} sequences are evaluated for makespan, and sequences with the lowest makesp are optimal.

A 4-job, 4-machine ordered flowshop problem will be solved to exemplify the above procedu. Processing times are shown in Table I. The matrix representing the processing time is called ordered matrix. Figure 1 shows the partial sequences formed at each stage of solution and the co pleted sequences. The optimal sequence is 2 4 3 1 with a makespan of 524.

TABLE I. *Ordered Matrix Problem*

Job	Machine			
	A	B	C	D
1	42	56	44	32
2	66	88	69	51
3	69	92	72	53
4	89	119	93	68

Further improvements in the computation time of the algorithm may be made by the followi modification. In Figure 1, we notice that the solution development is similar to that of the branch a bound method (although many branches considered in the general problem are eliminated). Thus computing suitable upper bounds on various nodes, it may be possible to eliminate many branche This idea has not yet been incorporated in the computer program that was developed. Such a procedu will limit the total program of sequences that need to be generated to 2^{n-1} or less, and hence the co putation times for problems may be reduced from the values above.

One characteristic of all the sequences generated by the algorithm is that once the position the largest job is fixed, all jobs preceding it are always arranged in an ascending order of processi

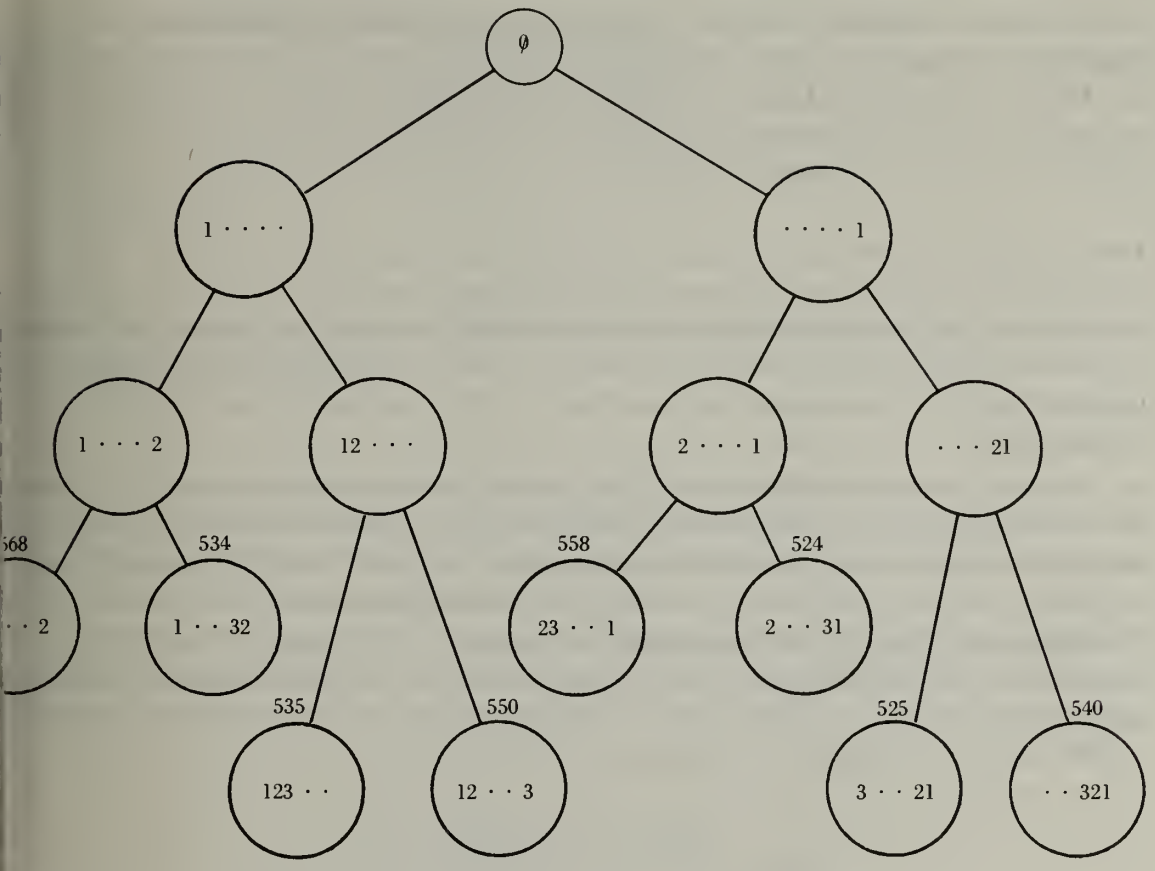


FIGURE 1. Graphical representation of the solution development.

, from the job in position 1, to the job appearing to the left of the largest job. Similarly, all jobs eding the largest job are arranged according to the descending order of processing times. A nce satisfying this characteristic (ascending order followed by descending order) will be con- ed as having a "pyramid" structure (see Appendix for further discussion).

he optimal sequence for an ordered flowshop problem, as described in the preceding section, imal only if we limit consideration to permutation schedules (schedules with the same order of n all machines). This can be seen by using the flowshop problem in Table II. This problem meets the requirements for ordered matrix problems and has sequence 1 2 as the optimal permutation ule with a makespan equal to 29; however, the general schedule given in Table III has a make-

TABLE II. *A Two-job Ordered Matrix Problem*

Job	Machine			
	A	B	C	D
1	8	6	4	7
2	5	2	1	4

TABLE III. *General Job Sequence for the Problem in Table II*

Machine	Sequence
A	1 2
B	1 2
C	2 1
D	2 1

span equal to 28. From this problem it can be concluded that there may be general schedules w lower total elapsed time than that of the optimal permutation schedule for ordered flowshop proble

DISCUSSION

Complete enumeration of an n-job problem with no passing involves generation of $n!$ sequenc The proposed algorithm, on the other hand, generates only 2^{n-1} sequences. Thus for a 10-job proble the complete enumeration and the proposed algorithm will generate 3,628,800 and 512 sequenc respectively. Since no other (general flowshop type) optimizing algorithms were available for solvi 10-job problems efficiently, no computational comparisons could be made. It was felt that proble with up to 15 jobs can be handled in a reasonable amount of time. As the computation time is det ministic, Table IV gives computation time for problems in each of the three categories as obtained an IBM 370/145. The algorithm was programmed in FORTRAN IV.

TABLE IV. *Computation Times*

Problem size $n \times m$	No. of sequences	Computation time (sec.)
10×20	512	11.5
10×50	512	27.5
15×20	16,384	535.7

The optimizing solution procedure for the flowshop ordered matrix problem is very efficie when compared with optimizing procedures for the general flowshop problem. The existing optimizi methods can efficiently solve problems involving relatively few jobs. In fact, it has been suggest by Karp [2] that these problems will remain intractable perpetually. Karp considers a solution pr cedure to be satisfactory if an algorithm terminates within a number of steps bounded by a polynomi in the length of the input. For the sequencing problem the length of the input is equivalent to t number of jobs. The proposed algorithm may not be satisfactory from this point of view; however, t ordered matrix procedure is a substantial improvement over all optimizing algorithms for the gener flowshop problem.

APPENDIX

Let N represent a set of n jobs and M represent a set of m machines. Let P_{ij} denote the processi time of job i on machine j.

Let N be partitioned into two mutually exclusive sets σ_1 and σ_2; σ_1 contains at least one job and nay be empty (we will assume without loss of generality that the job with the largest processing times lways included in σ_i). Theorem: For an n job m machine ordered matrix flowshop problem with and σ_2 as defined above, there exists an optimal makespan sequence of the form $\sigma_1\sigma_2$ where jobs r_1 are arranged in ascending order of processing times followed by jobs in σ_2 arranged in descending er of processing times.

To prove this theorem we will make use of theorems and proofs in [4]. Figure 2 gives a schematic resentation of a typical sequence S (using $n=15$ and $m=6$) where for convenience the job in ition i is labeled as job i. Let k denote the job with the largest processing times and t be the machine ı the largest processing times.

Mathematically, the two properties of the ordered flowshop problem stated earlier can be given as ows:

(i) If $P_{ir} > P_{jr}$ for $i, j \epsilon N$ and some $r \epsilon M$, then $P_{ix} \geq P_{jx}$ for all $x \epsilon M$.

(ii) If $P_{ir} > P_{ix}$ for some $i \epsilon N$ and $r, x \epsilon M$, then $P_{yr} \geq P_{yx}$ for all $y \epsilon N$.

h these properties we have $P_{kt} \geq P_{ij}$ for all $i \epsilon N$ and all $j \epsilon M$, where k and t represent the largest and machine (in terms of processing times), respectively.

Using contradictions and properties of the ordered matrix it is easy to show that the critical path always pass through P_{kt}. As an example, in Figure 1, $k=5$, $t=3$, and the solid lines represent a cal path giving the makespan.

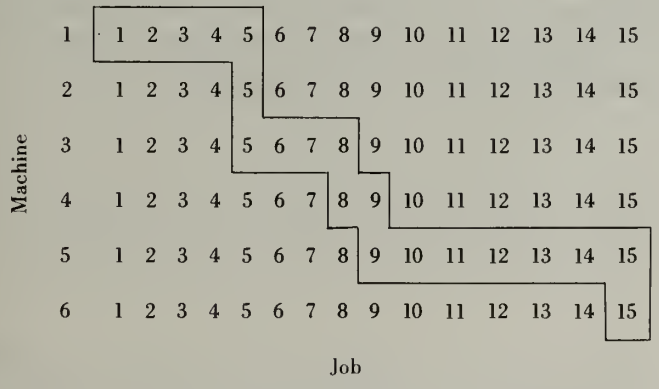


FIGURE 2. Schematic representation of makespan for sequence S.

Now consider a subproblem involving the first k jobs in sequence S and the first t machines. For subproblem the maximum processing time for all jobs occurs on the last machine (machine t). problem then can be solved by the algorithm described in [4] which states that the minimum espan sequence is given by arranging jobs in ascending order of processing times.

Returning to the original $N \times M$ problem, it can be seen that arranging jobs prior to job k in ascend- order will reduce the length of critical path (mathematically, it will not increase the length) and :e the makespan.

Using similar arguments, i.e., considering a subproblem which includes jobs $k, k+1, \ldots, n$ machines $t, t+1, \ldots, m$, one can show that arranging jobs in descending order will improve the espan.

Finally, consider two subsets of N given by σ_1 and σ_2 defined earlier. As mentioned earlier, let th largest job be included in σ_1. Let σ_1 include i jobs $(i=1, 2, \ldots, n)$. There are $\binom{n-1}{i-1}$ ways which one can choose jobs in σ_1. Hence there are $\binom{n-1}{i-1}$ sequences in which jobs in σ_1 can be a ranged in ascending order followed by jobs in σ_2 in descending order. In all, there are a total of

$$\sum_{i=1}^{n} \binom{n-1}{i-1} = 2^{n-1}$$

sequences. Note that each of these sequences has a pyramid structure. Also, the proposed algorith generates 2^{n-1} sequences having pyramid structures. From the theorem above, at least one optim sequence has the pyramid structure. Hence the algorithm will generate at least one optimum sequenc

REFERENCES

[1] Johnson, S. M., "Optimal Two and Three State Production Schedules with Set-up Times Included. Nav. Res. Log. Quart., *1*, 61–68, (1954).

[2] Karp, R. M., "Some Remarks on the Complexity of Combinatorial Optimization Problem," pr sented at the 41st Meeting of ORSA, New Orleans, Louisiana, Apr. 1972.

[3] Smith, M. L., R. A. Dudek, and S. S. Panwalkar, "Job Sequencing with Ordered Processing Tim Matrix," presented at the 43rd Meeting of ORSA, Milwaukee, Wisconsin, May 1973.

[4] Smith, M. L., S. S. Panwalkar, and R. A. Dudek, "Flowshop Sequencing Problem with Ordere Processing Time Matrices," Man. Sci., *21*, 544–549 (1975).

MEAN SQUARE INVARIANT FORECASTERS FOR THE WEIBULL DISTRIBUTION*

J. Tiago de Oliveira§

Faculty of Sciences
University of Lisbon
Lisbon, Portugal

and

Sebastian B. Littauer†

Columbia University
New York, N.Y.

ABSTRACT

Many techniques of forecasting are based upon extrapolation from time series. While such techniques have useful applications, they entail strong assumptions which are not explicitly enunciated. Furthermore, the time series approach is not based on an indigenous forecast principle. The first attack from the present point of view was initiated by S. S. Wilks.

Of particular interest over a wide range of operational situations in reliability, for example, is the behavior of the extremes of the Weibull and Gumbel distributions. Here we formulate forecasters for the minima of various forms of these distributions. The forecasters are determined for minimization in mean square of the distance. From n original observations the forecaster provides the minimum of the next m observations when the original distribution is maintained.

For each of the forecasters developed, tables of efficiency have been calculated and included in the appendix. An explicit example has been included for one of the forecasters. Its performance has been demonstrated by the use of Monte Carlo technique. The results indicate that the forecaster can be used in practice with satisfactory results.

INTRODUCTION

A basic scientific problem which has many ramifications is that of estimating future process outıes from an observed set of outcomes. This is a problem of long standing, that is, in essence, the lamental problem of forecasting. This problem can be, and has been, approached in many ways. it is not often recognized and expressed as a problem of forecasting. An early (the first) fundamental tribution to forecasting future process outcomes from a given set of, say, n process outcomes was le by S. S. Wilks [22, 23] in his pioneering papers on "Non-parametric tolerance intervals" in ch he introduced a new statistical interval, referred to in the literature either as a tolerance or a liction interval. It turns out that Wilks' results have great practical usefulness as well as simplicity elegance which are hard to match.

---

*Dedicated to the memory of E. J. Gumbel.

§Research of this author supported in part by the Calouste Gulbenkian Foundation and NSF Grant GP 8478 while at Columniversity.

†Research of this author supported in part by NSF Grant GP 8478.

Currently, much attention is being focused on the reliability of manufactured product as expresse in terms of product life. Very useful results in this connection have been obtained by Johns an Lieberman [9] for product life obeying the Weibull form of distribution. The authors did not, howeve express their results directly in the form of forecasting about some future outcomes of the proces based on, say, n previous observations.

Study of the particular problem of devising ways of giving measures of product life, which i important and challenging in itself, is perhaps more significant when approached from the point of vie of forecasting. Most forecasting effort has been directed to smoothing techniques when a forecas mode has been assumed—providing merely bounds of variation for an assumed forecast relation.

The present approach, while taking off from an assumed distribution function, provides as a for caster a function of $n+1, n+2, \ldots, n+m$ future possible process outcomes, based on the values o a set of n previous outcomes. The criterion for choice of the estimation-forecast function is the min mization of the mean square distance between the observed values and the ones forecast. This approac is quite different from current practice, and it is hoped that the present paper has some significance, bot operationally and theoretically, for the general scientific problem of forecasting as contrasted t smoothing.

1. THE PROBLEM

In a previous paper, Tiago de Oliveira [20], some results were derived about the direct computa tion, from observed values, of mean square forecasters and minimal average length forecasting regions These results are here applied to obtain mean square forecasters for Weibull and Gumbel distributions In a subsequent paper [21], lower bounds for the mean-square error of forecasting were obtained, thus providing a technique to evaluate the efficiency of the proposed forecasters. Here only a brief sketch will be given of the general results to be applied.

First, recall that the 3-parameter Weibull distribution function

$$(1) \qquad \begin{aligned} W(X;\lambda,\delta,\kappa) &= 0, && \text{if } x \leqslant \lambda \\ &= 1-\exp\left[-\left(\frac{X-\lambda}{\delta}\right)^{\kappa}\right], && \text{if } x \geqslant \lambda \end{aligned}$$

where λ is the location parameter, $\delta\ (>0)$ is the dispersion (location) parameter, and $\kappa\ (>0)$ is the shape parameter. If κ is known, $\kappa = \kappa_0$, considerable simplification results. Also, assuming the location parameter $\lambda = \lambda_0$, one obtains by the transformation $Y = -\ln(X-\lambda_0)$ the Gumbel largest values distribution

$$\exp\left(-\frac{Y-\xi}{\tau}\right)$$

where $\xi = -\ln\delta$ and $\tau = 1/\kappa$. Here ξ is the new location parameter and τ is the new dispersion parameter. We will consider only the cases where there are but two parameters, (λ, δ), $(\kappa = \kappa_0)$ or (ξ, τ), $(\lambda = \lambda_0)$, in the Weibull distribution or the Gumbel distribution, respectively.

2. THE GENERAL RESULTS ON QUASI-LINEAR INVARIANCE

Consider a sample of n observations $(x_1, x_2, \ldots, x_n)$ whose distribution is known about a location parameter λ and with respect to a dispersion parameter δ. Suppose further that a subsequent

ence $(x_{n+1}, \ldots, x_{n+m})$ from the same source is obtained, and that from the first sample it is ed to make a forecast on the behavior of a quasi-linearly invariant function

$$Z = \psi(X_{n+1}, \ldots, X_{n+m})$$

e second sample sequence, i.e., a function with the following property:

$$\psi(\lambda + \delta X_{n+1}, \ldots, \lambda + \delta X_{n+m}) = \lambda + \delta\psi(X_{n+1}, \ldots, X_{n+m}).$$

quasi-linear invariance property is needed for subsequent application.

Now consider the likelihood function of $(X_1, \ldots, X_n; Z)$ which, because of the quasi-linear invari- property of Z, can be expressed in the form

$$\frac{1}{\delta^{n+1}} \mathscr{L}\left(\frac{X_1 - \lambda}{\delta}, \ldots, \frac{X_n - \lambda}{\delta}; \frac{Z - \lambda}{\delta}\right).$$

Among the generally quasi-linearly invariant statistics Z are the (sample) average, median, largest (maximum), smallest value (minimum), and, more generally, linear functions of the order statistics. The present problem is concerned with finding a quasi-linearly invariant statistic

$$f(X_1, \ldots, X_n)$$

observed sample $(X_1, \ldots, X_n)$ which approaches optimally in the mean square sense to Z. The ion must therefore satisfy the following conditions:

$$f(\lambda + \delta X_1, \ldots, \lambda + \delta X_n) = \lambda + \delta f(X_1, \ldots, X_n),$$

$$E^2 = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [f(x_1, \ldots, x_n) - z]^2 \cdot \frac{1}{\delta^{n+1}} \mathscr{L}\left(\frac{x_1 - \lambda}{\delta}, \ldots, \frac{x_n - \lambda}{\delta}; \frac{z - \lambda}{\delta}\right) dx_1 \ldots dx_n dz$$

inimum.

We now consider expression (7) more in detail. On account of the quasi-linear invariance of $(X_{n+1}, \ldots, X_{n+m})$ by the linear transformation

$$X_i = \lambda + \delta X_i'$$

$$Z = \lambda + \delta Z',$$

tain

$$E^2 = \delta^2 \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [f(x_1', \ldots, x_n') - z']^2 \cdot \mathscr{L}(x_1', \ldots, x_n'; z')\, dx_1' \ldots dx_n' dz' = \delta^2 E_0^2$$

where E_0^2 results from E^2 when $\lambda=0$, $\delta=1$. We will return now to the use of (X, Z) even when deal with reduced values.

Note that E_0^2 and E^2 represent the *same mean square error*, expressed, however, in *different uni*

The general solution for the forecaster, as given in [20], is

(8)
$$f(x_1, \ldots, x_n)=\frac{\int_0^\infty d\beta\,\beta^{n-1}\int_{-\infty}^\infty d\alpha\,\overline{\mathscr{L}}(\alpha+\beta x_1, \ldots, \alpha+\beta x_n)\,[\mu_m(\alpha+\beta x_1, \ldots, \alpha+\beta x_n)-\alpha]}{\int_0^\infty d\beta\,\beta^n\int_{-\infty}^\infty d\alpha\,\overline{\mathscr{L}}(\alpha+\beta x_1, \ldots, \alpha+\beta x_n)}$$

where

(8′)
$$\overline{\mathscr{L}}(x_1, \ldots, x_n)=\int_{-\infty}^\infty dz\,\mathscr{L}(x_1, \ldots, x_n; z)$$

denotes the marginal likelihood of the sample, and

(9)
$$\mu_m(x_1, \ldots, x_n)=\int_{-\infty}^\infty dz\,z\,\mathscr{L}(x_1, \ldots, x_n; z)/\overline{\mathscr{L}}(x_1, \ldots, x_n)$$

denotes the conditional mean.

The value of E^2 can be split simply by setting $f-z=f-\mu_m-(z-\mu_m)$, which yields:

$$E_0^2=\int_{-\infty}^\infty \cdots \int_{-\infty}^\infty (z-\mu_m)^2\,\mathscr{L}(x_1, \ldots, x_n; z)dx_1 \cdots dx_n dz$$

(10)
$$+\int_{-\infty}^\infty \cdots \int_{-\infty}^\infty [f(x_1, \ldots, x_n)-\mu_m]^2\,\overline{\mathscr{L}}(x_1, \ldots, x_n)dx_1 \cdots d$$

$$=\sigma_{n,m}^2+D_{n,m}^2(f)$$

= variance of z about the conditional mean + mean square error of f (about the conditional mean). T function f minimizes the 2nd summand, since $\sigma_{n,m}^2$ is a constant. In the case of independence, the fi component is, exactly, the variance of z, because $\mu_m(x_1, \ldots, x_n)$ is a constant, as a consequence the splitting of $\mathscr{L}(x_1, \ldots, x_n; z)$ as $\overline{\mathscr{L}}(x_1, \ldots, x_n)h(z)=g(x_1)\cdots g(x_n)h(z)$. It was shown [21] that, in the general case, if we denote by $\varphi(\lambda, \delta)$ $(-\infty<\lambda<\infty,\ 0<\delta<\infty)$

$$\int_{-\infty}^\infty \cdots \int_{-\infty}^\infty \mu_m(\lambda+\delta x_1, \ldots, \lambda+\delta x_n)\,\overline{\mathscr{L}}(x_1, \ldots, x_n)dx_1 \cdots dx_n$$

and by $w_n(\lambda, \delta)$

$$\int_{-\infty}^\infty \cdots \int_{-\infty}^\infty \frac{\mathscr{L}^2(x_1, \ldots, x_n)}{\delta^n\,\mathscr{L}(\lambda+\delta x_1, \ldots, \lambda+\delta x_n)}\,dx_1 \cdots dx_n,$$

$$B^2 = \sup_{(\lambda,\delta)} \frac{(\lambda - \varphi(\lambda, \delta) + \varphi(0, 1)\delta)^2}{w_n(\lambda, \delta) - 2\delta + \delta^2}$$

lower bound for the mean-square error

$$D^2_{n,m}(f) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [f(x_1, \ldots, x_n) - \mu_m(x_1, \ldots, x_n)]^2 \overline{\mathscr{L}}(x_1, \ldots, x_n) dx_1 \ldots dx_n.$$

Note that B^2 is indeterminate for $\lambda = 0$, $\delta = 1$. For the case of independence, since $\varphi(\lambda, \delta) = \mu_m$ is stant and $w_n(\lambda, \delta) = w^n(\lambda, \delta)$ with

$$w(\lambda, \delta) = \int_{-\infty}^{\infty} \frac{g^2(x)}{\delta g(\lambda + \delta x)} dx,$$

wer bound has the simpler form

$$B^2 = \sup_{(\lambda,\delta)} \frac{(\lambda + \delta\mu_m - \mu_m)^2}{w^n(\lambda, \delta) - 2\delta + \delta^2}.$$

subevaluation of the efficiency (this lower bound is not necessarily sharp and, if sharp, may e attained) is then given as

$$B^2/D^2_{n,m}(f).$$

We will apply these results here, although, owing to computational difficulties, in some cases ill use a lower, but computable, bound $\overline{B}^2$. In case of regularity we can use a bound $B'^2 (\leq B^2)$. ound, analogous to the Cramér-Rao bounds for estimation, is given by

$$B'^2 \sim \frac{s_0\mu_m^2 - 2\, s_1\mu_m + s_2 - 1}{n\,(s_0 s_2 - s_1^2 - s_0)}$$

$$s_\alpha = \int_{-\infty}^{\infty} x^\alpha \frac{g'^2(x)}{g(x)} dx.$$

hus the subevaluation of the efficiency is given by

$$B'^2/D^2_{n,m}(f).$$

ubevaluation of the efficiency will be referred to merely as the efficiency.

3. THE SHAPE PARAMETER (κ_0) KNOWN IN WEIBULL DISTRIBUTION

The relations developed above will now be applied directly to the Weibull distribution when shape parameter is known. Suppose that we consider the quasi-linear function Z to be

$$Z = \min(X_{n+1}, \ldots, X_{n+m}). \tag{16}$$

Since we are dealing with independence, owing to the stability property of the Weibull distributi we have

$$\overline{\mathscr{L}}(x_1, \ldots, x_n) = W'(x_1; 0, 1, \kappa_0) \ldots W'(x_n; 0, 1, \kappa_0) \tag{17}$$

and

$$\mathscr{L}(x_1, \ldots, x_n; z) = \overline{\mathscr{L}}(x_1, \ldots, x_n) m^{1/\kappa_0} W'(m^{1/\kappa_0} z; 0, 1, \kappa_0), \tag{18}$$

where $W'(x; 0, 1, \kappa_0)$ denotes the density of $W(x; 0, 1, \kappa_0)$, that is, $W'(x; 0, 1, \kappa_0) = 0$ if $x \leq$ $W'(x; 0, 1, \kappa_0) = \kappa_0 x^{\kappa_0 - 1} e^{-(x^{\kappa_0})}$ if $x \geq 0$.

Consequently, we have

$$\begin{aligned} \mu_m &= \int_{-\infty}^{\infty} dz m^{1/\kappa_0} z W'(m^{1/\kappa_0} z; 0, 1, \kappa_0) = \Gamma(1 + 1/\kappa_0)/m^{1/\kappa_0} \\ &= \mu(\kappa_0)/m^{1/\kappa_0} \qquad (\mu(\kappa_0) = \mu_1). \end{aligned} \tag{19}$$

As $W'(x; 0, 1, \kappa_0) = 0$ if $x \leq 0$ the integration takes place in the region $\alpha + \beta x_i \geq 0$, that is, $\alpha \geq -\beta$ where $\ell = \min(x_1, \ldots, x_n)$.

The forecaster can then be written as $f(x_1, \ldots, x_n)$ where

$$f(x_1, \ldots, x_n) = \frac{\int_0^{\infty} d\beta \beta^{n-1} \int_{-\beta\ell}^{\infty} d\alpha (\mu_m - \alpha) \prod_{i=1}^{n} W'(\alpha + \beta x_i; 0, 1, \kappa_0)}{\int_0^{\infty} d\beta \beta^n \int_{-\beta\ell}^{\infty} d\alpha \prod_{i=1}^{n} W'(\alpha + \beta x_i; 0, 1, \kappa_0)}, \tag{20}$$

and, taking $\eta = \alpha + \beta\ell$,

$$f(x_1, \ldots, x_n) = \ell + \frac{\int_0^{\infty} d\beta \beta^{n-1} \int_0^{\infty} d\eta (\mu_m - \eta) \prod_{i=1}^{n} W'(\eta + \beta(x_i - \ell); 0, 1, \kappa_0)}{\int_0^{\infty} d\beta \beta^n \int_0^{\infty} d\eta \prod_{i=1}^{n} W'(\eta + \beta(x_i - \ell); 0, 1, \kappa_0)}. \tag{20'}$$

Computation facilities have no utility here because they would imply the computation of the forecast f for any n (sample size) and for any (practically) observable sample mean.

In the case of the exponential distribution ($\kappa_0 = 1$), we have

$$W'(X; 0, 1, 1) = 0, \qquad \text{if } X \leq 0$$
$$= e^{-x} \qquad \text{if } X \geq 0, \ \mu_m = 1/m,$$

we obtain

$$f(x_1, \ldots, x_n) = \ell + \left(\frac{1}{m} - \frac{1}{n}\right)[\bar{x} - \ell]$$

For other values of κ_0 the computation is cumbersome; as an example consider the Rayleigh dis-
ution where $\kappa_0 = 2$. Let us now determine a lower bound for efficiency computation. For B^2 we must
pute $w(\lambda, \delta)$. In many cases the supremum is attained at the indeterminacy point $(\lambda, \delta) = (0, 1)$.
sequently, to facilitate effective computability, we will take $\lambda = 0$, so that

$$w(0, \delta) = 1/[\delta^{\kappa_0}(2 - \delta^{\kappa_0})] \qquad \text{for } 0 < \delta^{\kappa_0} < 2,$$

e $g(X) = W'(X; 0, 1, \kappa_0)$.
From this we compute the lower bound

$$\bar{B}^2 = \lim_{\delta \to 1} \frac{\mu_m^2 (\delta - 1)^2}{\dfrac{1}{\delta^{\kappa_0 n}(2 - \delta^{\kappa_0})^n} - 2\delta + \delta^2} = \frac{\mu_m^2}{1 + n\kappa_0^2} \sim \frac{\mu_m^2}{n\kappa_0^2}.$$

evident that $B^{-2} \leq B^2$. The evaluation of the asymptotic efficiency is then given as

$$e_m(\kappa_0; f) = \lim_{n \to \infty} \frac{\mu_m^2}{n\kappa_0^2 D_{n,m}^2(f)} = \frac{\mu^2(\kappa_0)}{\kappa_0^2 m^{2/\kappa_0} \lim\limits_{n \to \infty} n D_{n,m}^2(f)}$$

proposed forecaster f.
For $\kappa_0 = 1$, $\bar{B}^2$ is asymptotically the value of $D_{n,m}^2(f)$ for the best forecaster, already given.
Since

$$s_\alpha = [(\kappa_0 - 1)^2 + \alpha(\kappa_0 + \alpha - 2)]\Gamma\left(1 + \frac{\alpha - 2}{\kappa_0}\right),$$

ound B'^2, for $\kappa_0 > 2$, is

$$B'^2 \sim \frac{\kappa_0^2 - 2\kappa_0(\kappa_0 - 1)\Gamma(1 - 1/\kappa_0)\mu_m + (\kappa_0 - 1)^2\Gamma(1 - 2/\kappa_0)\mu_m^2}{n\kappa_0^2(\kappa_0 - 1)^2[\Gamma(1 - 2/\kappa_0) - \Gamma^2(1 - 1/\kappa_0)]}$$

immediately evident that, asymptotically, $\bar{B}^2 \leq B'^2$.

4. A STUDY OF SOME FORECASTERS FOR KNOWN SHAPE PARAMETER ($\kappa_0 < 2$)

It seems desirable to explore different forecasters for different ranges of κ_0 because of comput tional difficulties. We will therefore examine the region $\kappa_0 \leq 2$, with a hint from the result for $\kappa_0 = 1$

The special simple form of the forecaster for the exponential distribution may suggest the use the function

$$\ell + a(\kappa_0)(\bar{x} - \ell) \tag{26}$$

as a forecaster, where $a(\kappa_0)$ is determined so as to minimize the mean square error

$$D^2_{n,m} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [\ell + a(\kappa_0)(\bar{x} - \ell) - \mu_m]^2 \overline{\mathscr{L}}(x_1, \ldots, x_n) dx_1 \ldots dx_n.$$

This suggestion is reinforced by the fact that the lower limit λ has ℓ as a sufficient statistic (ove estimator).

The sample average $\bar{X}$ is a quasi-linear statistic converging to $\mu(\kappa_0) = \Gamma^2(1 + 1/\kappa_0)$, with varianc

$$[\Gamma(1 + 2/\kappa_0) - \Gamma^2(1 + 1/\kappa_0)]/n = b(\kappa_0)/n.$$

ℓ has the expected value $\mu_n = \mu(\kappa_0)/n^{1/\kappa_0}$, its variance being

$$\frac{\Gamma(1 + 2/\kappa_0) - \Gamma^2(1 + 1/\kappa_0)}{n^{2/\kappa_0}} = b(\kappa_0)/n^{2/\kappa_0};$$

the correlation coefficient $\rho_n = \rho_n(\bar{x}, \ell)$ converges to zero (see Tiago de Oliveira [19] and Rosengaar [15]).

Thus, the mean-square error $D^2_{n,m}$, as a function of a, can be written as $E[(\ell + a(\bar{x} - \ell) - \mu_m)^2]$ $E[(\ell - \mu_m)^2] + 2E[(\ell - \mu_m)(\bar{x} - \ell)]a + E[(\bar{x} - \ell)^2]a^2$ where E denotes expected value. Using th decomposition $\ell - \mu_m = \ell - \mu_n + \mu_n - \mu_m$ and $\bar{x} - \ell = \bar{x} - \mu_1 + \mu_1 - \mu_n + \mu_n - \ell$, we have

$$E[(\ell - \mu_m)^2] = \frac{b(\kappa_0)}{n^{2/\kappa_0}} + (\mu_n - \mu_m)^2$$

$$E[(\ell - \mu_m)(\bar{x} - \ell)] = \rho_n \frac{b(\kappa_0)}{n^{1/\kappa_0 + 1/2}} - \frac{b(\kappa_0)}{n^{2/\kappa_0}} + (\mu_n - \mu_m)(\mu_1 - \mu_n)$$

$$E[(\bar{x} - \ell)^2] = \frac{b(\kappa_0)}{n} + (\mu_1 - \mu_n)^2 + \frac{b(\kappa_0)}{n^{2/\kappa_0}} - 2\rho_n \frac{b(\kappa_0)}{n^{1/\kappa_0 + 1/2}}.$$

The minimum of the quadratic is attained at some value whose limit is

$$a = \mu_n/\mu_1 = m^{-1/\kappa_0}, \tag{27}$$

nce the forecaster becomes: $\ell' = \ell + m^{-1/\kappa_0}(\bar{x} - \ell)$. The asymptotic values of the mean square r are then:

$$D^2_{n,m} \sim \frac{\mu_m^2 b(\kappa_0)}{\mu_1^2 n} = \frac{b(\kappa_0)}{n m^{2/\kappa_0}}, \qquad \kappa_0 < 2$$

$$\sim \frac{m - 2\sqrt{m} + 2 - \pi/4}{mn}, \qquad \kappa_0 = 2$$

$$\sim \frac{\Gamma\left(1 + \frac{2}{\kappa_0}\right)(\mu_1 - \mu_m)^2}{\mu_1^2 n^{2/\kappa_0}}, \qquad \kappa_0 > 2.$$

sequently, the asymptotic efficiency is, since $\bar{B}^2 \sim \frac{\mu_m^2}{n\kappa_0^2}$,

$$e_m(\kappa_0) = \frac{[\mu(\kappa_0)]^2}{\kappa_0^2 b(\kappa_0)} \qquad \kappa_0 < 2$$

$$= \frac{\pi}{16(m - 2\sqrt{m} + 2 - \pi/4)} \qquad \kappa_0 = 2$$

$$= 0 \qquad \kappa_0 > 2.$$

Zero efficiency for $\kappa_0 > 2$ results from the slow convergency of ℓ to the parameter λ, an effect in this region overcomes the more rapid convergence of the average $\bar{X}$. The efficiency for $\kappa_0 < 2$ dependent of m. A short table of the efficiency for the above forecaster is (for $m = 1$, only):

κ_0	1/3	1/2	1	2
$e(\kappa_0)$	0.5	0.8	1	0.91

forecaster for $1/2 < \kappa_0 < 2$ is reasonably efficient. Outside this range for the shape parameter, ıer forecaster must be found. A full set of values of this efficiency is given in Table 1 in the ›ndix.

Let us now consider the use of a sample quantile q and the smallest value ℓ. It is apparent (by rgument given above), that, since the variance of a quantile is of order $1/n$, a linear combination

$$\ell + a(q - \ell)$$

›e efficient only in the region $\kappa_0 \leq 2$; it will be useful in the region $\kappa_0 < 1/2$. For $\kappa = \kappa_0$ known, let note by ν the quantile of probability $\omega = W(\nu; 0, 1, \kappa_0)$. It is well known (Cramér [2]) that the

sample quantile q for probability ω is asymptotically normal with expected value ν and asympto variance of the order of $\omega(1-\omega)/nW'^2(\nu; 0, 1, \kappa_0)$.

Since in the present case we have

$$W(x; 0, 1, \kappa_0) = 1 - e^{-x^{\kappa_0}}, \quad x \geq 0$$

and

$$W'(x; 0, 1, \kappa_0) = \kappa_0 x^{\kappa_0 - 1} e^{-x^{\kappa_0}}, \quad x \geq 0,$$

the asymptotic variance is of the order of

$$\frac{e^{(\nu^{\kappa_0})} - 1}{\kappa_0^2 n \nu^{2(\kappa_0 - 1)}}.$$

The mean square error is

$$E[(\ell + a(q-\ell) - \mu_m)^2] = E[(\ell - \mu_m)^2] + 2E[(\ell - \mu_m)(q-\ell)]a + E[(q-\ell)^2]a^2,$$

with

$$E[(\ell - \mu_m)^2] = \frac{b(\kappa_0)}{n^{2/\kappa_0}} + (\mu_n - \mu_m)^2$$

$$E[(\ell - \mu_m)(q-\ell)] = \rho_n \tau_n \frac{b(\kappa_0)^{1/2}}{n^{1/\kappa_0}} - \frac{b(\kappa_0)}{n^{2/\kappa_0}} + (\mu_n - \mu_m)(\nu_n - \mu_n)$$

$$E[(q-\ell)^2] = \tau_n^2 + (\nu_n - \mu_n)^2 + \frac{b(\kappa_0)}{n^{2/\kappa_0}} - 2\rho_n \tau_n \frac{b(\kappa_0)^{1/2}}{n^{1/\kappa_0}},$$

where we have made a decomposition analogous to the previous one and where ν_n, τ_n^2, and ρ_n dend the exact expected value and the variance of the sample quantile and the correlation coefficient betwe the sample quantile and the minimum. As might be expected, $\rho_n \to 0$ (Rosengaard [15]).*

Denoting by ν the limiting value of ν_n, we see that the limiting value of a that minimizes the mea square error is

$$a = \frac{\mu_m}{\nu},$$

*In Cramér [2] it is shown that $(\nu_n - \nu_0) = o(n^{1/2})$.

the asymptotic mean-square error for $\kappa_0 < 2$ is

$$D^2_{n,m} \sim \frac{\mu_m^2 (e^{\nu^{\kappa_0}} - 1)}{n\kappa_0^2 \nu^{2\kappa_0}} .$$

For $\kappa_0 = 2$ we have

$$D^2_{n,m} \sim \frac{1}{m} \left\{ \frac{\pi(e^{\nu^2} - 1)}{16\nu^4} + \left(1 - \frac{\sqrt{\pi}}{2\sqrt{m}\nu}\right)^2 \right\}.$$

We will only consider the problem for $\kappa_0 < 2$. The case for $\kappa_0 = 2$ could be dealt with numerically. minimum mean-square error for variation of ν is attained for $\nu^{\kappa_0} = 1.593624$, and we have roximately

$$D^2_{n,m} \sim \frac{\mu_m^2}{n\kappa_0^2} \left(\frac{e^{1.593624} - 1}{(1.593624)^2} \right) = \frac{1.5441386 \mu_m^2}{n\kappa_0^2} ,$$

hat the asymptotic efficiency is

$$e_m(\kappa_0) = \frac{1}{1.5441386} = 0.65.$$

Taking the approximation $\nu^{\kappa_0} = 1.60$, the constant efficiency is $\frac{1}{1.5441532} = 0.65$ also. The forecaster be

$$\ell + \mu(\kappa_0)\,(1.60m)^{-1/\kappa_0}(q - \ell)$$

e q is then the sample quantile for probability $\omega = 1 - e^{-\nu^{\kappa_0}} = 1 - e^{-1.60} = 0.796$. This technique, e asymptotic efficiency is independent of m and κ_0, is to be used for values of $\kappa_0 \leq 0.40$ where the iency of the forecaster $\ell + a(\bar{x} - \ell)$ is smaller than 0.65. For $\kappa_0 = 2$ a better forecaster will be idered.

THE MAXIMUM LIKELIHOOD FORECASTER FOR THE SHAPE PARAMETER > 2) KNOWN

The maximum likelihood estimators of λ and δ for $\kappa_0 > 2$ known are given by the equations

$$\frac{\sum (x_i - \hat{\lambda})^{\kappa_0 - 1}}{\sum (x_i - \hat{\lambda})^{\kappa_0}} = \frac{\kappa_0 - 1}{n\kappa_0} \sum (x_i - \hat{\lambda})^{-1}$$

$$\hat{\delta}^{\kappa_0} = \frac{1}{n} \sum (x_i - \hat{\lambda})^{\kappa_0}$$

so that the forecaster of $\lambda+\mu_m\delta$ to be used must be $\hat{\lambda}+\mu_m\hat{\delta}$. Since, for $\lambda=0$, $\delta=1$, the variances a covariances are asymptotically

$$V(\hat{\lambda}) \sim \frac{\kappa_0^2}{n\Delta}, \quad V(\hat{\delta}) \sim \frac{(\kappa_0-1)^2\Gamma(1-2/\kappa_0)}{n\Delta},$$

$$C(\hat{\lambda}, \hat{\delta}) \sim \frac{-\kappa_0(\kappa_0-1)\Gamma(1-1/\kappa_0)}{n\Delta},$$

$$\Delta=\kappa_0^2(\kappa_0-1)^2[\Gamma(1-2/\kappa_0)-\Gamma^2(1-1/\kappa_0)],$$

the mean square error of the forecaster $\hat{\lambda}+\mu_m\hat{\delta}$ is

$$D^2_{m,\,m}(\hat{\lambda}+\mu_m\hat{\delta}) \sim \frac{\kappa_0^2-2\kappa_0(\kappa_0-1)\Gamma(1-1/\kappa_0)\mu_m+(\kappa_0-1)^2\Gamma(1-2/\kappa_0)\mu_m^2}{n\kappa_0^2(\kappa_0-1)^2[\Gamma(1-2/\kappa_0)-\Gamma^2(1-1/\kappa_0)]},$$

This is precisely equal to the asymptotic value of $(B')^2$, whence the efficiency is 1. This techniq requires the use of computers for determining $\hat{\lambda}$ and $\hat{\delta}$. Since, however, it is desirable to have simpl and more practical techniques where computer facilities are not available, we consider in the ne section some all range forecasters in this category. Such forecasters can be quite useful since t range $\kappa_0 \leq 2$ is most important for practical application.

6. SOME ALL-RANGE FORECASTERS

The difficulties found above in respect to computability and efficiency of some proposed forecaste may suggest the consideration of the traditional approach, using quantiles and moments, since the variances are of order n^{-1} at the lower bounds.

For example, consider the quantile q for the probability $\omega=W(\mu_m;\ 0,\ 1,\ \kappa_0)$, whose asymptoti efficiency for the bound $\bar{B}^2$ is given as

$$e_m(\kappa_0)=\frac{\mu_m^{2\kappa_0}}{e^{\mu_m^{\kappa_0}}-1}=\frac{[\mu(\kappa_0)]^{2\kappa_0}}{m^2[e^{\mu^{\kappa_0}(\kappa_0)}m-1]}, \tag{33}$$

which is approximated by $\mu_1^{\kappa_0}/m$ for large m. For $m=1$ (one-step forecast) we have the following tabl of efficiencies:

κ_0	0.25	0.5	1	2
$e_1(\kappa_0)$	0.60	0.64	0.58	0.52

This short table shows that the quantile $W(\mu_m;\ 0,\ 1,\ \kappa_0)$ whose efficiency fades out with m, is no useful in the range $\kappa_0 \leq 2$. Values of this efficiency are given in Table 2 of the Appendix.

For $\kappa_0 > 2$ the efficiency is given, using the bound B'^2, by

$$e_m(\kappa_0)=\frac{\mu_m^{2(\kappa_0-1)}(\kappa_0^2-2\kappa_0(\kappa_0-1)\Gamma(1-1/\kappa_0)\mu_m+(\kappa_0-1)^2\Gamma(1-2/\kappa_0)\mu_m^2)}{(e^{\mu_m^{\kappa_0}}-1)(\kappa_0-1)^2[\Gamma(1-2/\kappa_0)-\Gamma^2(1-1/\kappa_0)]} \tag{34}$$

$\mu_m = \mu(\kappa_0)/m^{1/\kappa_0} = \Gamma(1+1/\kappa_0)/m^{1/\kappa_0}$. The short table, for $m=1$ is:

κ_0	2.5	3.0	3.5	4.0	4.5	5.0
$e_1(\kappa_0)$	.53	.55	.57	.58	.59	.60

hows that the quantile forecaster has an efficiency of about 0.6. A full table of the efficiency for
> 2 is given in Table 3.

Consider now the use of a forecaster based on the very classical method of moments,

$$\bar{x}+as$$

re s denotes the standard deviation of the original sample. The mean square error is $D^2_{n,m} = E[(\bar{x}+as-\mu_m)^2] = E[(\bar{x}-\mu_m)^2]+2E[(\bar{x}-\mu_m)s]a+E(s^2)a^2$. The covariance between $\bar{x}$ and s, C_n, pproximated by $[\beta_1(\kappa_0)b(\kappa_0)]/2n$ (Gumbel and Carlson, [5]), where $\beta_1(\kappa_0)$ denotes the skewness fficient. With

$$\sigma_n(\kappa_0) \sim \sqrt{b(\kappa_0)}\left(1-\frac{\beta_2(\kappa_0)+3}{8n}\right)$$

expected value of s, where $\beta_2(\kappa_0)$ denotes the kurtosis coefficient (the fourth central moment over square of the variance), we have

$$D^2_{n,m} = \frac{b(\kappa_0)}{n}+(\mu_1-\mu_m)^2+2[C_n+(\mu_1-\mu_m)\sigma_n(\kappa_0)]a+b(\kappa_0)a^2.$$

The value at which the quadratic attains the minimum converges to

$$a = -\frac{\mu_1-\mu_m}{\sqrt{b(\kappa_0)}}$$

at the asymptotic value of the mean-square error is

$$D^2_{n,m} \sim \frac{b(\kappa_0)}{n}\left\{1-\frac{\beta_1(\kappa_0)(\mu_1-\mu_m)}{\sqrt{b(\kappa_0)}}+\frac{[\beta_2(\kappa_0)+3](\mu_1-\mu_m)^2}{4b(\kappa_0)}\right\}$$

e (Gumbel, [4]):

$$\beta_1(\kappa_0) = [\Gamma(1+3/\kappa_0)-3\Gamma(1+2/\kappa_0)\Gamma(1+1/\kappa_0)+2\Gamma^3(1+1/\kappa_0)]/b^{3/2}(\kappa_0)$$

$_0) = [\Gamma(1+4/\kappa_0)-4\Gamma(1+3/\kappa_0)\Gamma(1+1/\kappa_0)+6\Gamma(1+2/\kappa_0)\Gamma^2(1+1/\kappa_0)-3\Gamma^4(1+1/\kappa_0)]/b^2(\kappa_0)$.

The forecaster is then

$$\bar{x} - \left(\frac{\mu_1 - \mu_m}{b(\kappa_0)^{1/2}} \right) s. \tag{37}$$

Before continuing, a short proof of the asymptotic value of σ_n will be given. In Cramér [2] have the following relation,

$$E(s-\sigma_n)^2 = \frac{(\beta_2(\kappa_0)-1)b(\kappa_0)}{4n} + O(n^{-2}).$$

Developing the square and taking expected values, we have

$$\frac{n-1}{n} b(\kappa_0) - \sigma_n^2 = \frac{(\beta_2(\kappa_0)-1)b(\kappa_0)}{4n} + O(n^{-2})$$

from which the desired result follows.

The asymptotic efficiency with respect to $\bar{B}^2$ is

$$e_m(\kappa_0) = \frac{\mu_m^2}{\kappa_0^2 b(\kappa_0) \left\{ 1 - \frac{\beta_1(\kappa_0)(\mu_1-\mu_m)}{\sqrt{b(\kappa_0)}} + \frac{(\beta_2(\kappa_0)+3)(\mu_1-\mu_m)^2}{4b(\kappa_0)} \right\}}. \tag{38}$$

The asymptotic efficiency for $\kappa_0 > 2$, with respect to B'^2, is given by

(39)

$$e_m(\kappa_0) = \frac{\kappa_0^2 - 2\kappa_0(\kappa_0-1)\Gamma(1-1/\kappa_0)\mu_m + (\kappa_0-1)^2\Gamma(1-2/\kappa_0)\mu_m^2}{\kappa_0^2(\kappa_0-1)^2 b(\kappa_0)\left[\Gamma\left(1-\frac{2}{\kappa_0}\right) - \Gamma^2\left(1-\frac{1}{\kappa_0}\right)\right]\left\{1 - \frac{\beta_1(\kappa_0)(\mu_1-\mu_m)}{\sqrt{b(\kappa_0)}} + \frac{(\beta_2(\kappa_0)+3)(\mu_1-\mu_m)^2}{4b(\kappa_0)}\right.}$$

A short table for the first efficiency for $m=1(\mu_1=\mu_m)$ is:

κ_0	0.5	1.0	2.0
$e_1(\kappa_0)$	0.80	1.00	0.91

A table of the first efficiency for $\kappa_0 \leq 2$ is given in Table 4 of the Appendix; the second efficienc is given in Table 5 for $\kappa_0 > 2$. The second evaluation is closer to the true efficiency since $\bar{B}^2 \leq B'$ It readily may be shown that for the second efficiency at the limit $\kappa_0 \downarrow 2$, the value obtained is equ to that for the first efficiency at $\kappa_0 = 2$.

7. THE CASE FOR THE GUMBEL DISTRIBUTION

Let us consider now the Gumbel distribution, but for the sake of commodity, we will consider t maxima distribution $\exp(-e^{-Y})$. The minima are dealt with by symmetry. Computation of the fo

ter $f(Y_1, \ldots, Y_n)$ by the formula given in (8) gives

$$f(Y_1, \ldots, Y_n) = \frac{\int_0^\infty d\beta\beta^{n-1} \frac{e^{-\beta\Sigma Y_i}}{(\Sigma e^{-\beta Y_i})^n} \left(\gamma + \ln m + \frac{\Gamma'(n)}{\Gamma(n)} - \ln \Sigma e^{-\beta Y_i} \right)}{\int_0^\infty d\beta\beta^n \frac{e^{-\beta\Sigma Y_i}}{(\Sigma e^{-\beta Y_i})^n}}.$$

e formula is as uncomputable as the one given before. Since we will investigate other forecasters, will need the lower bound. This one can be computed by the formula for B'^2; its value is

$$B'^2 \sim \left[1 + \frac{6}{\pi^2} (1 + \ln m)^2 \right] / n.$$

$m = 1$ (one-step forecasting) its value is $1.61\ldots n$, as given in [21]; the value of μ_m is $\gamma + \ln m = 0.577$ m. As it is not natural to use here the smallest observed value, we will only consider as forecasters quantile for the point μ_m and the linear combination $\bar{y} + as$.

The various computations follow as before (using the same formulae) so that we will only give the ılts.

Quantile forecaster: its asymptotic mean square error is $D^2_{n,m} \sim e^{2\mu_m}(e^{e^{-\mu_m}} - 1)/n$ so that the ciency is

$$e_m = \frac{\left[1 + \frac{6}{\pi^2} (1 + \ln m)^2 \right]}{e^{2\gamma}(e^{e^{-\gamma/m}} - 1) m^2},$$

efficiency for $m = 1$ being $e_1 = 0.67$. Table 6 in the Appendix gives the values of this efficiency.

Moments forecaster: its asymptotic mean square error, where $\beta_1 = 1.14$, $\beta_2 = 5.4$, is

$$D^2_{n,m} \sim \frac{\pi^2}{6n} \left\{ 1 + \frac{\beta_1 \ln m}{\sqrt{\pi^2/6}} + \frac{(\beta_2 + 3)(\ln m)^2}{\pi^2/6} \right\}$$

at the efficiency is

$$e_m = \frac{\frac{6}{\pi^2} \left[1 + \frac{6}{\pi^2} (1 + \ln m)^2 \right]}{1 + \frac{\beta_1 \ln m}{\sqrt{\pi^2/6}} + \frac{(\beta_2 + 3)(\ln m)^2}{\pi^2/6}}.$$

The efficiency for $m = 1$ is $e_1 = 0.98$, the forecaster being

$$\bar{x} + \frac{\sqrt{6} \ln m}{\pi} s = \bar{x} + \frac{\ln m}{1.28} s.$$

Table 7 of the Appendix gives the values of these efficiencies.

Evidently the maximum likelihood forecaster (of the form $\hat{\lambda}+\mu_m\hat{\delta}$, see Gumbel [4]) has efficie one. In case the maximum likelihood forecaster is not used the quantile forecaster is preferable to moments forecaster for $m>1$.

8. THE LOCATION PARAMETER (λ_0) KNOWN IN THE WEIBULL DISTRIBUTION

In that case let us suppose now that λ has a value λ_o known; for simplicity take $\lambda_o=0$. Take th $Y_i=-\ln X_i$, $W=-\ln Z$. Their distribution is obviously, as said before, the Gumbel largest value d tribution with parameters $\xi=-\ln\delta$, $\tau=1/\kappa$. We may remark that if Z is any order statistic of $(X_n$ $\ldots, X_{n+m})$ then $W=-\ln Z$ is the complementary order statistic of $(-\ln X_{n+1}, \ldots, -\ln X_{n+m})$ ow to the monotonic decreasing character of the transformation.

Suppose now that for $(Y_1, \ldots, Y_n; W)$ we examine the forecaster $f(Y_1, \ldots, Y_n)$, verifying co ditions (6) and (7); that is,

$$f(\xi+\tau Y_1, \ldots, \xi+\tau Y_n)=\xi+\tau f(Y_1, \ldots, Y_n)$$

and

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}[w-f(y_1, \ldots, y_n)]^2\,\tilde{\mathscr{L}}(y_1, \ldots, y_n;\ w)\,dy_1\ldots dy_n dw,$$

a minimum, $\tilde{\mathscr{L}}(y_i, \ldots, y_n;\ w)$ denoting the likelihood function of $(y_1, \ldots, {}_n;\ w)$ as transform from the likelihood $\mathscr{L}(x_1, \ldots, x_n; z)$. Now putting $f(-\ln X_1, \ldots, -\ln X_n)=-\ln g(X_1, \ldots, X_n$ those relations are transformed to

$$g(\delta X_1^{\kappa}, \ldots, \delta X_n^{\kappa})=\delta g^{\kappa}(X_1, \ldots, X_n)$$

and

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\ln\frac{g(X_1, \ldots, X_n)}{z}\right)^2\mathscr{L}(X_1, \ldots, X_n; z)\,dX_1, \ldots, dX_n dz,$$

a minimum.

That shows that the technique of forecasting in the Gumbel distribution as applied to Y_i and does not give, effectively, the least squares forecaster of Z, but, say, the logarithmic least squa forecaster of Z.

9. EXAMPLE

Consider the forecaster $\ell'=\ell+m^{-1/\kappa_0}(\bar{x}-\ell)$ which derives from equations (26) and (27), wh in $W(X:\lambda, \delta, \kappa)$, λ is set equal to zero, $\delta=1$, and $\kappa=\kappa_0$, $0<\kappa_0\leq 2$, whence $W(X:0, 1, \kappa_0)=1-e^{-x}$.

In the forecaster equation, ℓ' is the forecast minimum from n observations on random varia X; $\bar{x}$ is the linear average of these n random values of X; ℓ is the minimum of $W(X:0, 1, \kappa_0)$ result

the first n values of X; m subsequent values of $W(X: 0, 1, \kappa_0)$ are obtained on a new set of m ran- values of X; $\bar{x}_m$ is the linear average of these m values. The forecast minimum ℓ' is compared the m observed values of $W(X: 0, 1, \kappa_0)$.

The results are given below: For

$\kappa_0 = 1.2$: $n = 50, m = 20, 40$
$\kappa_0 = 1.2$: $n = 100, m = 50$
$\kappa_0 = 1.2$: $n = 500, m = 50, 100, 200, 300, 400$
$\kappa_0 = 1.5$: $n = 50, m = 20, 40$
$\kappa_0 = 1.5$: $n = 100, m = 50$
$\kappa_0 = 1.5$: $n = 500, m = 50, 100, 200, 300, 400$
$\kappa_0 = 2.0$: $n = 50, m = 20, 40$
$\kappa_0 = 2.0$: $n = 100, m = 50$
$\kappa_0 = 2.0$: $n = 500, m = 100, 200, 300, 400.$

Two conclusions are apparent from these results. For each value of κ_0, the number among the ure observations which are smaller than the forecast minimum ℓ' decreases as n increases, as be expected. And the forecaster is more efficient for smaller values of κ_0, when n is large. The ncy is much poorer for $\kappa_0 = 2.0$. It seems that an effective range for the combination (κ_0, n) is found empirically.

The results contained below were obtained by Monte Carlo simulation using an H.P. 9810 pro- nable calculator. The values $\ell'_1, \ell'_2, \ldots$, denote the observed smallest values among the d sample of m, which are smaller than ℓ'.

We wish to express our appreciation to Professor Tuncel M. Yegulalp for his assistance in pro- the computations for determining the following data:

n	m	$\bar{x}_n$	$\bar{x}_m$	ℓ	ℓ'	ℓ_1'	ℓ_2'	ℓ_3'	ℓ_4'	ℓ_5'	ℓ_6'
50	20	1.106482	0.618368	0.087202	0.171168	0.013614	0.067171	0.090015	0.170916		
50	40	0.999691	0.864131	0.037770	0.082242	0.012288	0.026883	0.027323	0.057124		
100	50	0.993607	0.870766	0.058743	0.094630	0.009625	0.075735	0.089076			
500	50	0.955362	1.053420	0.017990	0.053974	$\geq \ell'$					
500	100	0.986448	0.811142	0.005847	0.026973	0.009693	0.021219				
500	200	0.948326	0.995920	0.009625	0.020975	$\geq \ell'$					
500	300	0.995920	0.995920	0.009625	0.016349	$\geq \ell'$					
500	400	0.938055	0.927183	0.000687	0.007048	0.004685					
50	20	0.802585	1.157086	0.041248	0.097344	$\geq \ell'$					
50	40	0.864939	0.919777	0.091047	0.157814	0.020942					
100	50	0.952770	0.885156	0.080250	0.144538	0.041757					
500	50	0.936661	0.987915	0.032320	0.098952	0.083899					
500	100	0.902707	0.901279	0.024205	0.065282	0.044599	0.059561				
500	200	0.931957	0.893814	0.009254	0.036226	0.010214					
500	300	0.906875	0.894213	0.003972	0.024120	$\geq \ell'$					
500	400	0.948098	0.958196	0.004700	0.022078	0.001939	0.018320				
50	20	0.936665	0.870119	0.254734	0.407218	0.138128	0.288591	0.386233			
50	40	0.818688	0.914252	0.089966	0.205187	0.018976	0.090137	0.102450			
100	50	0.822177	0.981068	0.100348	0.202430	$\geq \ell'$					
500	100	0.895189	0.837045	0.056219	0.140116	0.021859	0.023695	0.056533	0.063959	0.099370	0.118113
500	200	0.918345	0.862408	0.076011	0.135573	0.010172	0.045231	0.054819	0.064132	0.093357	0.097550
500	300	0.901596	0.894982	0.066431	0.114653	0.065947	0.076226	0.085539	0.095738	0.099727	
500	400	0.897788	0.884023	0.047098	0.089632	0.031186	0.046646	0.055452	0.076625	0.086074	

REFERENCES

[1] Chew, Victor, "Simultaneous Prediction Intervals," Technometrics *10*, 323–31 (1968).

[2] Cramer, H., *Mathematical Methods of Statistics* (Princeton University Press, 1946).

[3] David, H. A., *Order Statistics* (John Wiley & Sons, Inc., New York, 1970).

[4] Gumbel, E. J., *Statistics of Extremes* (Columbia University Press, 1958).

[5] Gumbel, E. J. and P. G. Carlson, "On the Asymptotic Covariance of the Sample Mean and St ard Deviation," Metron *XVIII* (1956).

[6] Hahn, C. J., "Factors for Calculating Two-Sided Prediction Intervals for Samples From a No Distribution," Journal of the American Statistical Association *64*, 878–88 (1969).

[7] Hahn, D. J., "Additional Factors for Calculating Prediction Intervals for Samples From a No Population," Journal of the American Statistical Association *65*, 1668–76 (1970).

[8] Harter, H. L., *Order Statistics and Their Use in Testing and Estimation*, Vol. 2, U.S. Governr Printing Office (Washington, D.C. 1969).

[9] Johns, M. V. and G. J. Lieberman, "An Exact Asymptotically Efficient Confidence Bounc Reliability in the Case of Weibull Distributions," Technometrics *8*, (1966).

[10] Lieberman, G. J., "Prediction Regions for Several Predictions From a Single Regression Li Technometrics *3*, 409–17 (1961).

[11] Mann, Nancy R., "Warranty Periods Based on Three Ordered Sample Observations fro Weibull Population," I.E.E.E. Transactions on Reliability *R–19*, 167–71 (1970).

[12] Mann, Nancy R., and Frank E. Grubbs, "Chi-Square Approximations for Exponential Paramet Prediction Intervals and Beta Percentiles," Journal of the American Statistical Associat *69*, 654–661 (1974).

[13] Mann, Nancy R., and Sam C. Saunders, "On Evaluation of Warranty Assurance When Life a Weibull Distribution," Biometrika *56*, 615–25 (1969).

[14] Mann, Nancy R., Ray E. Schafer, and Nozer D. Singpurwalla, *Methods for Statistical Anal of Reliability and Life Data* (John Wiley & Sons, Inc., New York, 1974).

[15] Rosengaard, A., Contributions a l'Etude des Liaisons Litites entre Differents Statistique d Echantillon, *These, Institutede Statistique*, Univ. Paris (1966).

[16] Sarhan, Ahmed E. and Bernard G. Greenberg, *Contributions to Order Statistics* (John Wiley & S Inc., New York, 1962).

[17] Saunders, Sam C., "On the Determination of a Safe Life For Distribution Classified by Fai Rate," Technometrics *10*, 361–77 (1968).

[18] Teichroew, D., "Tables of Expected Values of Order Statistics from Samples of Size 20 and l From the Normal Distribution," Annals of Mathematical Statistics *27*, 410–26 (1956).

[19] Tiago de Oliveira, J., "The Asymptotic Independence of the Mean and the Extreme Valu Univ. Lisboa, Rev. Fac. Ciencias 2(*A*), 8 (1962).

[20] Tiago de Oliveira, J., "Quasi-Linearly Invariant Prediction," Annals of Mathematical Statist *37* (1966).

[21] Tiago de Oliveira, J., "Efficiency Evaluations for Quasi-Linearly Invariant Predictors," Rev Belge de Statistique et Recherche Operationelle *9* (1968).

[22] Wilks, S., "Determination of Sample Sizes for Setting Tolerance Limits," Annals of Mathemat Statistics *12* (1941).

Wilks, S., "Statistical Prediction With Special Reference to the Problem of Tolerance Limits," Annals of Mathematical Statistics *13*, (1942).

APPENDIX

TABLES OF EFFICIENCIES

TABLE 1. Computations based on

$$e_m(\kappa_0) = [\mu(\kappa_0)]^2/\kappa_0^2\, b(\kappa_0)$$

κ_0	0.10000	0.20000	0.30000	0.40000	0.50000	0.60000	0.70000	0.80000	0.90000	1.00000
$e_m\kappa_0$	0.00054	0.09960	0.37996	0.63355	0.80000	0.89873	0.95424	0.98339	0.99655	1.00000
κ_0	1.10000	1.20000	1.30000	1.40000	1.50000	1.60000	1.70000	1.80000	1.90000	2.00000
$e_m\kappa_0$	0.99753	0.99149	0.98334	0.97401	0.96409	0.95396	0.94385	0.93393	0.92428	0.91495

TABLE 2. Computations based on

$$e_m(\kappa_0) = [\mu(\kappa_0)]^{2\kappa_0}/m^2\, [e^{\mu^{\kappa_0}(\kappa_0)/m} - 1]$$

κ_0	0.10000	0.20000	0.30000	0.40000	0.50000	0.60000	0.70000	0.80000	0.90000	1.00000
m										
1	0.22380	0.54150	0.63075	0.64753	0.64242	0.63072	0.61755	0.60471	0.59280	0.58198
2	0.59447	0.63339	0.57570	0.52518	0.48633	0.45638	0.43282	0.41387	0.39833	0.38537
3	0.64651	0.54524	0.46145	0.40667	0.36899	0.34164	0.32091	0.30467	0.29160	0.28086
4	0.60970	0.46206	0.37826	0.32798	0.29473	0.27113	0.25352	0.23986	0.22896	0.22005
5	0.55665	0.39703	0.31886	0.27388	0.24473	0.22429	0.20915	0.19749	0.18822	0.18067
6	0.50543	0.34673	0.27502	0.23478	0.20901	0.19109	0.17788	0.16774	0.15970	0.15316
7	0.46009	0.30720	0.24155	0.20531	0.18231	0.16638	0.15469	0.14573	0.13865	0.13290
8	0.42090	0.27550	0.21523	0.18236	0.16161	0.14730	0.13683	0.12881	0.12248	0.11735
9	0.38716	0.24959	0.19402	0.16398	0.14511	0.13213	0.12265	0.11540	0.10968	0.10505
10	0.35804	0.22805	0.17659	0.14895	0.13166	0.11979	0.11112	0.10451	0.09930	0.09508
15	0.25863	0.15903	0.12172	0.10208	0.08991	0.08161	0.07558	0.07099	0.06738	0.06447
20	0.20177	0.12196	0.09282	0.07761	0.06824	0.06187	0.05725	0.05374	0.05098	0.04876
25	0.16524	0.09887	0.07499	0.06260	0.05488	0.04982	0.04607	0.04323	0.04100	0.03921
30	0.13985	0.08312	0.06290	0.05245	0.04604	0.04169	0.03855	0.03616	0.03429	0.03278
35	0.12120	0.07170	0.05417	0.04513	0.03959	0.03585	0.03313	0.03108	0.02946	0.02817
40	0.10693	0.06303	0.04757	0.03961	0.03473	0.03144	0.02905	0.02725	0.02583	0.02469
45	0.09566	0.05623	0.04240	0.03529	0.03094	0.02799	0.02587	0.02426	0.02299	0.02198
50	0.08653	0.05076	0.03824	0.03181	0.02789	0.02523	0.02331	0.02186	0.02072	0.01980
60	0.07267	0.04248	0.03197	0.02658	0.02329	0.02107	0.01946	0.01825	0.01730	0.01653
70	0.06263	0.03653	0.02747	0.02283	0.02000	0.01809	0.01671	0.01566	0.01484	0.01418
80	0.05502	0.03204	0.02408	0.02001	0.01752	0.01585	0.01463	0.01372	0.01300	0.01242
90	0.04906	0.02853	0.02143	0.01780	0.01559	0.01410	0.01302	0.01220	0.01156	0.01105
100	0.04427	0.02571	0.01931	0.01604	0.01404	0.01270	0.01172	0.01099	0.01041	0.00995

TABLE 2. Computations based on (Cont.)

$$e_m(\kappa_0) = [\mu(\kappa_0)]^{2\kappa_0}/m^2\ [e^{\mu^{\kappa_0}(\kappa_0)/m} - 1]$$

κ_0	1.10000	1.20000	1.30000	1.40000	1.50000	1.60000	1.70000	1.80000	1.90000	2.00000
m										
1	0.57221	0.56341	0.55548	0.54832	0.54183	0.53593	0.53055	0.52562	0.52110	0.51694
2	0.37441	0.36501	0.35686	0.34974	0.34345	0.33787	0.33288	0.32838	0.32432	0.32063
3	0.27187	0.26424	0.25768	0.25198	0.24698	0.24256	0.23862	0.23509	0.23191	0.22902
4	0.21264	0.20637	0.20099	0.19634	0.19226	0.18867	0.18547	0.18261	0.18003	0.17770
5	0.17440	0.16911	0.16459	0.16067	0.15725	0.15424	0.15156	0.14917	0.14701	0.14507
6	0.14775	0.14319	0.13929	0.13592	0.13298	0.13039	0.12809	0.12604	0.12419	0.12252
7	0.12814	0.12413	0.12071	0.11776	0.11518	0.11291	0.11090	0.10910	0.10748	0.10602
8	0.11311	0.10954	0.10649	0.10386	0.10157	0.09955	0.09776	0.09617	0.09473	0.09343
9	0.10123	0.09801	0.09526	0.09290	0.09083	0.08902	0.08741	0.08597	0.08468	0.08351
10	0.09160	0.08867	0.08618	0.08402	0.08215	0.08050	0.07903	0.07773	0.07655	0.07550
15	0.06207	0.06005	0.05833	0.05685	0.05556	0.05443	0.05343	0.05253	0.05173	0.05100
20	0.04693	0.04539	0.04408	0.04295	0.04197	0.04111	0.04035	0.03967	0.03906	0.03850
25	0.03772	0.03648	0.03543	0.03452	0.03372	0.03303	0.03241	0.03186	0.03137	0.03093
30	0.03154	0.03050	0.02961	0.02885	0.02818	0.02760	0.02708	0.02662	0.02621	0.02584
35	0.02710	0.02620	0.02544	0.02478	0.02421	0.02371	0.02326	0.02287	0.02251	0.02219
40	0.02375	0.02296	0.02229	0.02172	0.02121	0.02077	0.02038	0.02004	0.01972	0.01944
45	0.02114	0.02044	0.01984	0.01933	0.01888	0.01849	0.01814	0.01783	0.01755	0.01730
50	0.01904	0.01841	0.01787	0.01741	0.01701	0.01665	0.01634	0.01606	0.01581	0.01559
60	0.01590	0.01537	0.01492	0.01453	0.01419	0.01390	0.01364	0.01340	0.01319	0.01300
70	0.01364	0.01319	0.01280	0.01247	0.01218	0.01192	0.01170	0.01150	0.01132	0.01116
80	0.01195	0.01155	0.01121	0.01092	0.01066	0.01044	0.01024	0.01007	0.00991	0.00977
90	0.01063	0.01027	0.00997	0.00971	0.00948	0.00929	0.00911	0.00896	0.00881	0.00869
100	0.00957	0.00925	0.00898	0.00874	0.00854	0.00836	0.00820	0.00806	0.00794	0.00782

TABLE 3. Computations based on

$$e_m(\kappa_0) = \frac{\mu_m^{2(\kappa_0-1)}\,(\kappa_0^2 - 2\kappa_0\,(\kappa_0-1)\,\Gamma(1-1/\kappa_0)\mu_m + (\kappa_0-1)^2\,\Gamma(1-2/\kappa_0)\mu_m^2)}{(e^{\mu_m^{\kappa_0}}-1)\,(\kappa_0-1)^2\,[\Gamma(1-2/\kappa_0) - \Gamma^2(1-1/\kappa_0)]}$$

κ_0	2.10000	2.30000	2.40000	2.60000	2.90000	3.00000	3.30000	3.60000	4.00000	4.40000	4.70000	5.00000	6.00000	7.00000	8.00000	9.00000	10.00000
m																	
1	0.51930	0.52533	0.52871	0.53576	0.54628	0.54966	0.55921	0.56783	0.57788	0.58647	0.59208	0.59709	0.61038	0.61992	0.62704	0.63256	0.63693
2	0.34794	0.39467	0.41461	0.44887	0.48909	0.50015	0.52795	0.54950	0.57133	0.58767	0.59734	0.60536	0.62404	0.63533	0.64267	0.64769	0.65131
3	0.27497	0.34883	0.37859	0.42728	0.48049	0.49436	0.52765	0.55181	0.57466	0.59053	0.59934	0.60627	0.62080	0.62813	0.63206	0.63418	0.63536
4	0.23681	0.32835	0.36395	0.42042	0.47927	0.49402	0.52828	0.55191	0.57297	0.58659	0.59367	0.59889	0.60836	0.61166	0.61241	0.61199	0.61112
5	0.21390	0.31770	0.35700	0.41790	0.47902	0.49387	0.52739	0.54945	0.56802	0.57913	0.58443	0.58801	0.59289	0.59278	0.59094	0.58851	0.58605
6	0.19888	0.31158	0.35335	0.41682	0.47854	0.49313	0.52521	0.54540	0.56140	0.57012	0.57380	0.57590	0.57686	0.57393	0.57002	0.56602	0.56233
7	0.18839	0.30782	0.35128	0.41623	0.47764	0.49180	0.52217	0.54045	0.55400	0.56054	0.56278	0.56361	0.56122	0.55598	0.55040	0.54515	0.54049
8	0.18074	0.30539	0.35002	0.41576	0.47637	0.49002	0.51861	0.53504	0.54631	0.55090	0.55188	0.55159	0.54637	0.53920	0.53225	0.52600	0.52057
9	0.17494	0.30374	0.34921	0.41529	0.47482	0.48792	0.51473	0.52942	0.53861	0.54145	0.54131	0.54005	0.53240	0.52362	0.51555	0.50847	0.50242
10	0.17044	0.30258	0.34864	0.41476	0.47305	0.48560	0.51069	0.52374	0.53104	0.53231	0.53119	0.52908	0.51932	0.50918	0.50016	0.49241	0.48585
15	0.15794	0.29995	0.34696	0.41120	0.46275	0.47276	0.49042	0.49684	0.49680	0.49223	0.48754	0.48242	0.46545	0.45082	0.43883	0.42899	0.42093
20	0.15252	0.29887	0.34550	0.40677	0.45214	0.46009	0.47207	0.47375	0.46874	0.46043	0.45354	0.44660	0.42553	0.40851	0.39505	0.38425	0.37552
25	0.14966	0.29806	0.34387	0.40214	0.44221	0.44849	0.45605	0.45417	0.44559	0.43468	0.42631	0.41819	0.39453	0.37614	0.36188	0.35062	0.34159
30	0.14798	0.29727	0.34213	0.39758	0.43312	0.43801	0.44201	0.43739	0.42612	0.41331	0.40388	0.39494	0.36957	0.35035	0.33566	0.32414	0.31500
35	0.14690	0.29647	0.34036	0.39323	0.42484	0.42857	0.42964	0.42280	0.40942	0.39519	0.38499	0.37545	0.34890	0.32914	0.31421	0.30263	0.29344
40	0.14618	0.29563	0.33860	0.38908	0.41726	0.42001	0.41861	0.40996	0.39490	0.37954	0.36875	0.35877	0.33140	0.31131	0.29628	0.28467	0.27551
45	0.14566	0.29479	0.33685	0.38515	0.41032	0.41221	0.40871	0.39854	0.38209	0.36584	0.35458	0.34427	0.31631	0.29603	0.28096	0.26938	0.26030
50	0.14529	0.29395	0.33515	0.38146	0.40391	0.40507	0.39973	0.38827	0.37069	0.35371	0.34207	0.33150	0.30311	0.28275	0.26770	0.25619	0.24717
60	0.14478	0.29230	0.33187	0.37467	0.39248	0.39236	0.38402	0.37049	0.35110	0.33304	0.32088	0.30997	0.28106	0.26066	0.24573	0.23439	0.22558
70	0.14445	0.29066	0.32883	0.36859	0.38255	0.38141	0.37066	0.35555	0.33480	0.31598	0.30349	0.29232	0.26319	0.24289	0.22815	0.21704	0.20841
80	0.14424	0.28912	0.32597	0.36306	0.37378	0.37177	0.35909	0.34269	0.32095	0.30155	0.28882	0.27754	0.24834	0.22822	0.21370	0.20279	0.19437
90	0.14408	0.28762	0.32333	0.35801	0.36592	0.36322	0.34890	0.33146	0.30889	0.28913	0.27622	0.26489	0.23573	0.21579	0.20150	0.19082	0.18257
100	0.14394	0.28620	0.32081	0.35341	0.35885	0.35552	0.33983	0.32154	0.29833	0.27826	0.26528	0.25386	0.22484	0.20511	0.19107	0.18058	0.17255

TABLE 4. Computations based on

$$e_m(\kappa_0) = \frac{\mu_m^2}{\kappa_0^2\, b(\kappa_0) \left\{ 1 - \dfrac{\beta_1(\kappa_0)\ (\mu_1 - \mu_m)}{\sqrt{b\ (\kappa_0)}} + \dfrac{(\beta_2(\kappa_0) + 3)\ (\mu_1 - \mu_m)^2}{b\ (\kappa_0)} \right\}}$$

κ_0	0.10000	0.20000	0.30000	0.40000	0.50000	0.60000	0.70000	0.80000	0.90000	1.00000
m										
1	0.00054	0.09960	0.37996	0.63355	0.80000	0.89873	0.95424	0.98339	0.99655	1.00000
2	0.00000	0.00000	0.00000	0.00010	0.00116	0.00560	0.01626	0.03414	0.05755	0.08333
3	0.00000	0.00000	0.00000	0.00001	0.00016	0.00096	0.00324	0.00762	0.01414	0.02222
4	0.00000	0.00000	0.00000	0.00000	0.00005	0.00032	0.00120	0.00306	0.00605	0.01000
5	0.00000	0.00000	0.00000	0.00000	0.00002	0.00014	0.00058	0.00158	0.00329	0.00565
6	0.00000	0.00000	0.00000	0.00000	0.00001	0.00007	0.00033	0.00094	0.00204	0.00362
7	0.00000	0.00000	0.00000	0.00000	0.00000	0.00004	0.00021	0.00062	0.00138	0.00252
8	0.00000	0.00000	0.00000	0.00000	0.00000	0.00003	0.00014	0.00043	0.00099	0.00185
9	0.00000	0.00000	0.00000	0.00000	0.00000	0.00002	0.00010	0.00031	0.00074	0.00142
10	0.00000	0.00000	0.00000	0.00000	0.00000	0.00001	0.00007	0.00024	0.00057	0.00112
20	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00001	0.00004	0.00011	0.00025
30	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00001	0.00004	0.00011
50	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00001	0.00004
100										

κ_0	1.10000	1.20000	1.30000	1.40000	1.50000	1.60000	1.70000	1.80000	1.90000	2.00000
m										
1	0.99753	0.99149	0.98334	0.97401	0.96409	0.95396	0.94385	0.93393	0.92428	0.91495
2	0.10856	0.13142	0.15119	0.16783	0.18166	0.19307	0.20248	0.21025	0.21669	0.22205
3	0.03108	0.04001	0.04854	0.05640	0.06351	0.06985	0.07548	0.08047	0.08490	0.08882
4	0.01459	0.01946	0.02434	0.02904	0.03345	0.03754	0.04128	0.04470	0.04781	0.05064
5	0.00851	0.01166	0.01492	0.01814	0.02125	0.02419	0.02695	0.02951	0.03189	0.03408
6	0.00560	0.00785	0.01022	0.01263	0.01499	0.01727	0.01943	0.02147	0.02339	0.02518
7	0.00398	0.00568	0.00752	0.00941	0.01130	0.01314	0.01492	0.01661	0.01821	0.01973
8	0.00299	0.00433	0.00581	0.00735	0.00891	0.01046	0.01196	0.01340	0.01478	0.01609
9	0.00233	0.00342	0.00465	0.00594	0.00727	0.00859	0.00989	0.01115	0.01236	0.01352
10	0.00187	0.00279	0.00382	0.00493	0.00608	0.00724	0.00838	0.00949	0.01057	0.01161
20	0.00047	0.00076	0.00112	0.00154	0.00201	0.00250	0.00301	0.00353	0.00405	0.00458
30	0.00021	0.00037	0.00057	0.00081	0.00109	0.00139	0.00172	0.00206	0.00241	0.00277
50	0.00008	0.00015	0.00025	0.00037	0.00052	0.00069	0.00087	0.00108	0.00129	0.00151
100	0.00002	0.00005	0.00008	0.00013	0.00019	0.00027	0.00036	0.00046	0.00057	0.00066

TABLE 5. Computations based on

$$e_m(\kappa_0) = \frac{\kappa_0^2 - 2\kappa_0(\kappa_0 - 1)\,\Gamma(1 - 1/\kappa_0)\mu_m + (\kappa_0 - 1)^2\,\Gamma(1 - 2/\kappa_0)\mu_m^2}{\kappa_0^2\,(\kappa_0 - 1)^2 b(\kappa_0)\left[\Gamma\left(1 - \frac{2}{\kappa_0}\right) - \Gamma^2\left(1 - \frac{1}{\kappa_0}\right)\right]\left\{1 - \frac{\beta_1(\kappa_0)\,(\mu_1 - \mu_m)}{\sqrt{b(\kappa_0)}} + \frac{(\beta_2(\kappa_0) + 3)\,(\mu_1 - \mu_m)^2}{b(\kappa_0)}\right\}}$$

κ_0	2.10000	2.20000	2.30000	2.40000	2.50000	2.60000	2.70000	2.80000	2.90000	3.00000
m										
1	0.91696	0.91963	0.92273	0.92610	0.92961	0.93315	0.93667	0.94012	0.94347	0.94668
2	0.24844	0.27307	0.29596	0.31716	0.33677	0.35488	0.37162	0.38708	0.40138	0.41461
3	0.11212	0.13402	0.15448	0.17353	0.19123	0.20765	0.22288	0.23701	0.25012	0.26229
4	0.07178	0.09168	0.11030	0.12766	0.14380	0.15878	0.17270	0.18561	0.19761	0.20876
5	0.05391	0.07257	0.09003	0.10630	0.12144	0.13549	0.14854	0.16065	0.17190	0.18236
6	0.04413	0.06196	0.07864	0.09418	0.10862	0.12203	0.13448	0.14603	0.15676	0.16674
7	0.03806	0.05530	0.07141	0.08642	0.10036	0.11330	0.12530	0.13645	0.14679	0.15641
8	0.03396	0.05075	0.06644	0.08104	0.09460	0.10718	0.11885	0.12967	0.13972	0.14906
9	0.03103	0.04747	0.06282	0.07710	0.09035	0.10265	0.11405	0.12462	0.13443	0.14355
10	0.02883	0.04498	0.06006	0.07409	0.08710	0.09916	0.11034	0.12070	0.13032	0.13925
15	0.02297	0.03826	0.05250	0.06571	0.07795	0.08928	0.09976	0.10947	0.11847	0.12682
20	0.02043	0.03526	0.04904	0.06182	0.07364	0.08457	0.09467	0.10402	0.11268	0.12070
25	0.01902	0.03356	0.04705	0.05955	0.07110	0.08177	0.09163	0.10074	0.10917	0.11699
30	0.01813	0.03245	0.04575	0.05805	0.06941	0.07990	0.08958	0.09852	0.10679	0.11445
35	0.01751	0.03168	0.04482	0.05698	0.06819	0.07854	0.08808	0.09690	0.10505	0.11259
40	0.01705	0.03111	0.04413	0.05616	0.06727	0.07750	0.08694	0.09566	0.10371	0.11116
45	0.01670	0.03066	0.04359	0.05553	0.06654	0.07668	0.08604	0.09467	0.10264	0.11001
50	0.01643	0.03031	0.04315	0.05501	0.06594	0.07601	0.08530	0.09386	0.10176	0.10907
60	0.01602	0.02977	0.04249	0.05423	0.06503	0.07499	0.08415	0.09260	0.10040	0.10761
70	0.01573	0.02939	0.04201	0.05365	0.06436	0.07422	0.08330	0.09166	0.09938	0.10651
80	0.01551	0.02909	0.04164	0.05321	0.06385	0.07363	0.08264	0.09093	0.09858	0.10564
90	0.01534	0.02887	0.04135	0.05285	0.06343	0.07316	0.08210	0.09034	0.09793	0.10495
100	0.01521	0.02868	0.04111	0.05256	0.06309	0.07276	0.08166	0.08985	0.09740	0.10436
	3.10000	3.20000	3.30000	3.40000	3.50000	3.60000	3.70000	3.80000	3.90000	4.00000
	0.94977	0.95270	0.95548	0.95812	0.96060	0.96294	0.96514	0.96721	0.96916	0.97098
	0.42687	0.43822	0.44877	0.45857	0.46769	0.47620	0.48413	0.49155	0.49849	0.50496
	0.27361	0.28413	0.29395	0.30309	0.31164	0.31963	0.32712	0.33414	0.34074	0.34692
	0.21914	0.22880	0.23781	0.24623	0.25409	0.26146	0.26837	0.27486	0.28097	0.28669
	0.19210	0.20116	0.20963	0.21753	0.22492	0.23184	0.23834	0.24444	0.25018	0.25557
	0.17602	0.18466	0.19273	0.20027	0.20731	0.21392	0.22011	0.22593	0.23141	0.23655
	0.16535	0.17368	0.18146	0.18872	0.19550	0.20187	0.20783	0.21344	0.21872	0.22368
	0.15774	0.16583	0.17337	0.18042	0.18700	0.19318	0.19897	0.20441	0.20953	0.21434
	0.15202	0.15991	0.16727	0.17415	0.18057	0.18659	0.19224	0.19754	0.20254	0.20723
	0.14756	0.15529	0.16250	0.16922	0.17551	0.18141	0.18694	0.19213	0.19702	0.20161
	0.13457	0.14178	0.14850	0.15477	0.16063	0.16611	0.17125	0.17608	0.18063	0.18489
	0.12815	0.13507	0.14152	0.14753	0.15314	0.15840	0.16332	0.16794	0.17229	0.17637
	0.12423	0.13096	0.13723	0.14307	0.14852	0.15362	0.15840	0.16288	0.16710	0.17105
	0.12155	0.12815	0.13428	0.14000	0.14533	0.15032	0.15499	0.15938	0.16350	0.16736
	0.11958	0.12607	0.13211	0.13773	0.14297	0.14787	0.15246	0.15677	0.16082	0.16461
	0.11806	0.12447	0.13042	0.13597	0.14114	0.14597	0.15050	0.15474	0.15873	0.16247
	0.11684	0.12318	0.12907	0.13455	0.13966	0.14444	0.14891	0.15310	0.15704	0.16074
	0.11584	0.12212	0.12795	0.13338	0.13844	0.14317	0.14760	0.15175	0.15564	0.15930
	0.11428	0.12046	0.12621	0.13155	0.13653	0.14118	0.14553	0.14961	0.15344	0.15704
	0.11310	0.11921	0.12489	0.13017	0.13508	0.13968	0.14397	0.14800	0.15177	0.15532
	0.11218	0.11823	0.12385	0.12908	0.13394	0.13849	0.14273	0.14672	0.15045	0.15395
	0.11143	0.11743	0.12301	0.12819	0.13301	0.13751	0.14172	0.14567	0.14937	0.15284
	0.11080	0.11677	0.12230	0.12745	0.13223	0.13670	0.14088	0.14479	0.14846	0.15190

TABLE 5. Computations based on (Cont.)

$$e_m(\kappa_0) = \frac{\kappa_0^2 - 2\kappa_0(\kappa_0 - 1)\,\Gamma(1 - 1/\kappa_0)\mu_m + (\kappa_0 - 1)^2\,\Gamma(1 - 2/\kappa_0)\mu_m^2}{\kappa_0^2\,(\kappa_0 - 1)^2 b(\kappa_0)\left[\Gamma\left(1 - \frac{2}{\kappa_0}\right) - \Gamma^2\left(1 - \frac{1}{\kappa_0}\right)\right]\left\{1 - \frac{\beta_1(\kappa_0)\,(\mu_1 - \mu_m)}{\sqrt{b(\kappa_0)}} + \frac{(\beta_2(\kappa_0) + 3)\,(\mu_1 - \mu_m)^2}{b(\kappa_0)}\right\}}$$

κ_0	4.10000	4.20000	4.30000	4.40000	4.50000	4.60000	4.70000	4.80000	4.90000	5.00000
m										
1	0.97269	0.97431	0.97583	0.97722	0.97855	0.97978	0.98094	0.98202	0.98304	0.98399
2	0.51107	0.51680	0.52219	0.52725	0.53204	0.53656	0.54081	0.54485	0.54865	0.55227
3	0.35277	0.35828	0.36347	0.36838	0.37303	0.37744	0.38161	0.38558	0.38934	0.39293
4	0.29211	0.29722	0.30205	0.30662	0.31096	0.31507	0.31896	0.32268	0.32619	0.32956
5	0.26067	0.26549	0.27004	0.27435	0.27844	0.28232	0.28600	0.28951	0.29284	0.29603
6	0.24143	0.24602	0.25037	0.25448	0.25839	0.26210	0.26562	0.26898	0.27216	0.27521
7	0.22837	0.23280	0.23699	0.24096	0.24473	0.24830	0.25169	0.25493	0.25800	0.26095
8	0.21889	0.22319	0.22726	0.23110	0.23476	0.23823	0.24152	0.24467	0.24765	0.25051
9	0.21167	0.21586	0.21982	0.22357	0.22714	0.23053	0.23373	0.23680	0.23971	0.24250
10	0.20596	0.21006	0.21394	0.21761	0.22110	0.22441	0.22755	0.23056	0.23340	0.23613
15	0.18893	0.19274	0.19634	0.19975	0.20299	0.20607	0.20898	0.21177	0.21441	0.21694
20	0.18023	0.18387	0.18731	0.19057	0.19367	0.19661	0.19939	0.20205	0.20457	0.20699
25	0.17480	0.17833	0.18166	0.18482	0.18782	0.19067	0.19336	0.19594	0.19838	0.20073
30	0.17102	0.17447	0.17772	0.18081	0.18374	0.18651	0.18915	0.19166	0.19405	0.19633
35	0.16820	0.17159	0.17479	0.17781	0.18069	0.18341	0.18599	0.18846	0.19080	0.19304
40	0.16600	0.16934	0.17249	0.17547	0.17830	0.18098	0.18352	0.18595	0.18825	0.19046
45	0.16423	0.16752	0.17063	0.17357	0.17636	0.17901	0.18152	0.18392	0.18619	0.18837
50	0.16275	0.16601	0.16908	0.17199	0.17475	0.17737	0.17985	0.18222	0.18447	0.18662
60	0.16043	0.16363	0.16665	0.16950	0.17222	0.17478	0.17722	0.17955	0.18175	0.18386
70	0.15866	0.16182	0.16479	0.16761	0.17028	0.17281	0.17521	0.17750	0.17967	0.18175
80	0.15726	0.16038	0.16332	0.16610	0.16874	0.17124	0.17361	0.17587	0.17802	0.18007
90	0.15611	0.15920	0.16211	0.16486	0.16748	0.16995	0.17230	0.17454	0.17666	0.17869
100	0.15515	0.15821	0.16110	0.16383	0.16642	0.16887	0.17119	0.17341	0.17552	0.17753

κ_0	5.50000	6.00000	6.50000	7.00000	7.50000	8.00000	8.50000	9.00000	9.50000	10.00000
m										
1	0.98794	0.99079	0.99281	0.99424	0.99525	0.99591	0.99645	0.99664	0.99679	0.99682
2	0.56780	0.58004	0.58982	0.59782	0.60436	0.60987	0.61460	0.61846	0.62216	0.62531
3	0.40851	0.42101	0.43121	0.43971	0.44681	0.45289	0.45820	0.46267	0.46689	0.47058
4	0.34423	0.35607	0.36580	0.37396	0.38082	0.38673	0.39191	0.39632	0.40048	0.40413
5	0.30993	0.32120	0.33048	0.33828	0.34488	0.35057	0.35557	0.35985	0.36388	0.36743
6	0.28853	0.29933	0.30825	0.31577	0.32213	0.32764	0.33248	0.33663	0.34054	0.34399
7	0.27381	0.28425	0.29288	0.30016	0.30634	0.31168	0.31639	0.32043	0.32424	0.32761
8	0.26300	0.27314	0.28154	0.28863	0.29465	0.29986	0.30446	0.30841	0.31213	0.31542
9	0.25468	0.26459	0.27279	0.27972	0.28560	0.29071	0.29520	0.29908	0.30273	0.30596
10	0.24806	0.25776	0.26580	0.27259	0.27836	0.28337	0.28778	0.29159	0.29518	0.29835
15	0.22802	0.23704	0.24451	0.25084	0.25623	0.26091	0.26504	0.26861	0.27198	0.27496
20	0.21758	0.22619	0.23333	0.23938	0.24454	0.24902	0.25298	0.25641	0.25964	0.26250
25	0.21097	0.21931	0.22623	0.23209	0.23709	0.24143	0.24527	0.24860	0.25173	0.25452
30	0.20633	0.21447	0.22122	0.22694	0.23182	0.23606	0.23981	0.24306	0.24613	0.24885
35	0.20285	0.21082	0.21745	0.22306	0.22784	0.23201	0.23569	0.23888	0.24189	0.24456
40	0.20011	0.20796	0.21448	0.22000	0.22471	0.22880	0.23243	0.23557	0.23854	0.24117
45	0.19789	0.20563	0.21206	0.21750	0.22215	0.22619	0.22977	0.23287	0.23580	0.23840
50	0.19603	0.20368	0.21004	0.21542	0.22001	0.22401	0.22755	0.23061	0.23351	0.23608
60	0.19309	0.20060	0.20683	0.21211	0.21661	0.22053	0.22400	0.22701	0.22985	0.23238
70	0.19084	0.19823	0.20436	0.20956	0.21399	0.21786	0.22127	0.22424	0.22704	0.22953
80	0.18905	0.19634	0.20239	0.20752	0.21190	0.21571	0.21909	0.22202	0.22478	0.22724
90	0.18757	0.19478	0-20077	0.20584	0.21017	0.21394	0.21728	0.22018	0.22292	0.22535
100	0.18632	0.19347	0.19940	0.20443	0.20871	0.21245	0.21576	0.21863	0.22134	0.22375

TABLE 6. Computations based on

$$e_m = \frac{\left[1 + \frac{6}{\pi^2}(1 + \ln m)^2\right]}{e^{2\gamma}(e^{e^{-\gamma}/m} - 1)m^2},$$

1	2	3	4	5	6	7	8	9	10	15	20
0.67294	0.66695	0.62584	0.58335	0.54532	0.51222	0.48346	0.45831	0.43617	0.41652	0.34379	0.29635
25	30	35	40	45	50	60	70	80	90	100	
0.26249	0.23687	0.21666	0.20023	0.18655	0.17495	0.15626	0.14177	0.13015	0.12057	0.11253	

TABLE 7. Computations based on

$$e_m = \frac{\frac{6}{\pi^2}\left[1 + \frac{6}{\pi^2}(1 + \ln m)^2\right]}{1 + \frac{\beta_1 \ln m}{\sqrt{\pi^2/6}} + \frac{(\beta_2 + 3) \cdot (\ln m)^2}{\pi^2/6}}$$

1	2	3	4	5	6	7	8	9	10	15	20
0.97750	0.40975	0.27466	0.22518	0.19955	0.18373	0.17291	0.16497	0.15887	0.15401	0.13926	0.13151
25	30	35	40	45	50	60	70	80	90	100	
0.12658	0.12310	0.12047	0.11839	0.11670	0.11528	0.11302	0.11128	0.10989	0.10874	0.10776	

A COMPARISON OF SEVERAL ATTRIBUTE SAMPLING PLANS

Alan E. Gelfand
University of Connecticut
Storrs, Connecticut

ABSTRACT

The purpose of this research is to examine several types of procedures for attribute sampling inspection—the widely used Military Standard 105D plans [8], the lesser known Double Zero plans as developed by Ellis [4] and the Narrow Limit gaging plans of Ott and Mundel [9]. Each of the procedures is described with an effort made to illuminate their more subtle features. Then the plans are compared, whence it is revealed that (i) Narrow Limit gaging plans have a serious weakness in comparison to the others and (ii) Double Zero plans tend to be essentially conservative, but that sufficiently tight Military Standard 105D plans can be selected to achieve comparable performance in all ways.

TRODUCTION

he need for, and advantages of, employing a quality control program to guide a production ss is well established. One basic device for checking the stability of such a repetitive process statistical control chart initially developed by Shewart [10]. A sample of items is selected, and ch item a measurement on a process variable of interest is taken. These measurements are ned in conjunction with previous information on the process in order to make a decision as to er the process is in a state of control or is out of control.

n many situations the process variable is dichotomized and items are classified only in terms of ier they are "good" or "defective." The acceptability of an item may be determined in various For instance, we may have to refer to prescribed upper and lower drawing (or specification) , or we may note whether or not it withstands certain testing procedures, or we may just "eyeball" ier or not it is suitable for use. A sample of items is selected and examined to record the number ectives obtained. Such an approach is referred to as sampling inspection by attributes. Specifica- f a sample size to be inspected along with the maximum allowable number of defectives specifies ribute sampling plan. While such a "go—no go" approach retains less information than one has he actual measurement value for each item, it has its advantages in terms of facility, efficiency, ility, and cost considerations.

is not our intent to examine the question of assigning individual component tolerances. Develop- ch specifications is critical in attempting to achieve a desired overall level of performance for a r system composed of various components. A good survey discussion of such ideas is contained paper of Evans [5]. Rather, we shall assume that criteria for the acceptability of an item (compo- r system) have been determined and that lots of production items are being subjected to attribute ing inspection according to these criteria. Our concern focuses on the selection of an appropriate ute sampling plan and the ramifications of such a choice. In discussing attribute sampling plans ety of performance characteristics are often used. Among these, are the OC (operating char-

acteristic) curve, the AQL (acceptable quality level), the AOQL (acceptable outgoing quality leve and the LTPD (lot tolerance percent defective). Numerous texts discuss these concepts. Among th most well known are Burr [2], Grant [6], Brownlee [1], and Johnson and Leone [7]. Virtually all th information regarding the performance of an attribute sampling plan is contained in its OC curv The first extensive tabulation of such plans was developed by Dodge and Romig [3]. The emphasis i recent years has been on developing plans to achieve certain specified types of performance. Amon these are the extensive and widely used MIL STD 105D plans [8], the Ott-Mundel Narrow Limit Gagin plans [9], and the Double Zero plans developed by Ellis [4]. It is our intent to examine and compar these plans.

The format of this paper, then, is the following. In the second section we shall describe the plan In the third section we show that, despite their different origins and intents, the plans are directl mathematically comparable in terms of operating characteristic curves. We shall finally compare th plans according to:

1. Inherent assumptions and suitability for application.
2. Philosophy of lot acceptance.
3. Consumer and producer protection.
4. Relative cost.

2. A DESCRIPTION OF THE PLANS

The clearest description of any attribute sampling plan can be obtained by examining its OC curv which is a plot of P_a vs. p where P_a is the probability of acceptance of a lot containing a proportion, p of defectives. We let N indicate the lot size, and denote an attribute sampling plan as an (n, c) pla where n is the sample size to be inspected and c is the acceptance number. A lot is accepted if, at mos c defectives are found in the sample of n. At any value of p, P_a is calculated exactly using the hype geometric distribution by

$$P_a(p) = \frac{1}{\binom{N}{n}} \sum_{j=0}^{c} \binom{Np}{j} \binom{N(1-p)}{n-j}.$$

When N is large the hypergeometric distribution is usually approximated by the binomial distributio whence P_a becomes

$$P_a(p) \approx \sum_{j=0}^{c} \binom{n}{j} p^j (1-p)^{n-j}.$$

This expression is exact if we assume an infinite lot size. Moreover, if n is quite large and p i small, the Poisson distribution is employed to approximate P_a, yielding

$$P_a(p) \approx \sum_{j=0}^{c} \frac{(np)^j e^{-np}}{j!}.$$

Clearly, P_a decreases in p and $P_a(0) = 1$, $P_a(1) = 0$.

he attribute sampling plans described in Military Standard 105D are the most widely used in nited States today. These procedures, by their own description, have been developed for applica- requiring the "continuous" inspection of end items, components, and raw materials, as well as enance, administrative and data-keeping operations. These schemes are straightforward sampling which are tabulated such that, given the lot size, N, and a selected AQL from a set of preferred s, a particular (n, c) plan is specified. The AQL is nothing more than a choice of p which is deemed table. Specifying N and an AQL does not uniquely determine n and c, but rather they are deter- such that the domain for the probability of acceptance, P_a, at the AQL is usually somewhere en 0.88 and 0.99. The plans allow for tightened and reduced inspection as well, changing P_a priately. More precisely, tightened inspection decreases P_a at each p, hence lowering the OC Similarly, reduced inspection increases P_a at each p, thus raising the OC curve.

he Ott-Mundel Narrow Limit gaging plans and the Ellis Double Zero plans have been much less sed. This is due, in part, to the fact that neither of these approaches has been as extensively as the Military Standard 105D plans. Additionally, these plans require the determination of and lower specification limits for the process variable as opposed to the simple go—no go (de- —nondefective) determination needed for the Military Standard 105D plans. Finally, in terms respective critical characteristics which form the bases of these plans, we shall show that there exists a Military Standard 105D plan which will achieve comparable performance. These re- are not intended to disparage Narrow Limit gaging or Double Zero plans. In applications where rformance criteria which these plans are designed to control are crucial these approaches directly y plans which will provide the desired protection.

arrow Limit gaging plans and Double Zero plans are quite similar in their presumptions. Thus t surprising that Ott and Mundel discuss an application of their plans to a machine shop, while idicates his plans are appropriate for aircraft and metal working industries with the likelihood lar suitable areas of application.

et us describe the Double Zero sampling plans first. The thrust behind such plans is to reject ich are not zero-defective, i.e., totally free of defects. Sampling procedures which select an r some variant thereof, implicitly assume that the consumer is willing to accept lots containing entage of defective items. Clearly, zero-defect plans would be very useful in certain industrial tions.

ie following assumptions are inherent in so-called Double Zero sampling plans.

Double Zero plans never knowingly accept a lot that contains defective items. That is, only zero acceptance numbers are allowed.

In examining an item (or characteristic of an item) to determine whether or not it is defective, it is crucial to assume that the measured variable follows a normal distribution. Such an assumption is far from unreasonable. For example, experimentation in the metal-working industry has shown that, for over 95 percent of samples, measurements were obtained which formed an essentially normal frequency distribution.

In making the decision as to whether or not an item (or characteristic of an item) is defective, it is presumed that there is a region of prescribed dimensional tolerance characterized by D_L, the lower tolerance (or drawing) limit, and D_U, the upper tolerance (or drawing) limit. Additionally, the existence of a "gray area" is assumed, which extends below D_L to the lower suitable-for-use limit, S_L, and above D_U to the upper suitable-for-use limit S_U. The situation is described

in Figure 1. This "gray area" is established, not with the intent of accepting lots having ite within this area, but rather with the intent of insuring that no lots with items beyond this a will be accepted.

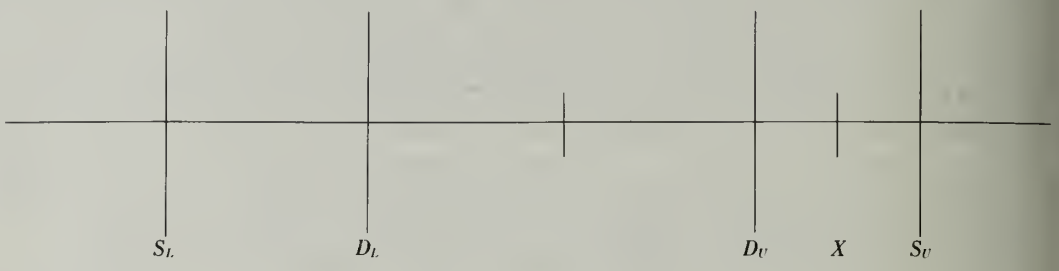


FIGURE 1

The distinguishing feature of Double Zero plans is in the examination of the extent of nonconfor ance of defective items rather than the percentage of nonconforming items. If, in Figure 1, an it yielded an observed measurement X, then the extent of nonconformance could be computed, a percent of the given tolerance, by

$$\frac{X - D_U}{D_U - D_L} \times 100.$$

For applications where product may still be suitable for use even if nonconforming, this ext is more useful than just a decision to reject. Note that if X fell between D_L and D_U, the observatio within specification and the notion of "extent of nonconformance" isn't meaningful.

A Double Zero sampling plan is characterized by specifying a percent of tolerance such th under the "worst conditions" (to be clarified shortly), the probability is essentially zero that an it in the lot will exceed this percent of tolerance. Moreover, in the terminology "Double Zero," while t first zero refers to the fact that only zero acceptance numbers are allowed, the second zero refers the fact that the above probability is claimed to be always essentially zero. More specifically, a typi Double Zero sampling plan might be denoted by 0050, and is intended to mean that a zero acceptar number is being used and the chance of there being an item in the lot which is defective by more th 50 percent of tolerance is virtually zero.

In determining what percent of tolerance to use for a particular item, one would choose a perce at most

$$\frac{S_U - D_U}{D_U - D_L} \times 100,$$

in order to be essentially certain that the item is, at worst, in the gray area. For example, if the diame of an item is to be within 0.002″ of specification and is suitable for use within 0.003″ of specificati then it is clear that any plan at least as tight as 0025 would be appropriate.

Let us elaborate further on the so-called "worst condition" mentioned above. Suppose an infin lot size with proportion of defectives p. Then, under assumption (2), the measurements will be inc

ıdent with common normal distribution such that $1-p$ of the area is between D_L and D_U, and p of area is outside. Now suppose we choose α $(0 \leq \alpha \leq 1)$ such that αp is the probability of an observa- ı above D_U and $(1-\alpha)p$ is the probability of an observation below D_L, and consider all possible mal distributions (means denoted by μ and variances denoted by σ^2) satisfying these requirements. en, if z_γ denotes the γ^{th} percentile under the standard (or unit) normal curve, $D_U = \mu + z_{1-\alpha p}\sigma$ and $= \mu + z_{(1-\alpha)p}\sigma$. Moreover, for any normal curve, the probability is essentially one that an observa- ı will fall between $\mu - 3\sigma$ and $\mu + 3\sigma$. Hence the extent of nonconformance can be considered as extent above D_U, i.e., $(\mu + 3\sigma) - D_U$ or the extent below D_L, i.e., $D_L - (\mu - 3\sigma)$. By expressing the ent as a percent of the total tolerance, we would obtain the percent above

$$\frac{(\mu + 3\sigma) - (\mu + z_{1-\alpha p}\sigma)}{(\mu + z_{1-\alpha p}\sigma) - \mu + z_{(1-\alpha)p}\sigma} \times 100 = \frac{3 - z_{1-\alpha p}}{z_{1-\alpha p} - z_{(1-\alpha)p}} \times 100 \equiv E(p;\, \alpha).$$

Note that the extent $E(p;\, \alpha)$ is independent of μ and σ can be calculated for any given p and α ng a standard normal curve. Due to symmetry considerations we need only examine the percent of rance above in terms of maximizing the extent. In other words, given any value of p, the "worst ıation," or condition of maximum extent, is achieved by obtaining

$$E(p) \equiv \max_{0 \leq \alpha \leq 1} E(p;\, \alpha).$$

We note that:

(1) Again $E(p)$ can be calculated for each p and is scale and location invariant, i.e., free of μ and σ^2.

(2) Clearly $E(p)$ is strictly increasing in p since $E(p;\alpha)$ increases in p at a fixed α. Thus $E(p)$ is a one-to-one function of p with $E(0) = 0$, $E(1) = \infty$.

(3) Perhaps surprisingly, in calculating $E(p)$, the maximum doesn't necessarily occur at $\alpha = 1$. For example, with $p = 0.06$ it is easy to calculate that $E(0.06,\, 5/6) = 33.6$ while $E(0.06,\, 1) = 31.6$ and in fact that $E(0.06) = 34.12$ and occurs at $\alpha = 0.8267$.

(4) The critical use of the normality assumption is apparent.

Since $E(p)$ is monotone in p, one can find p such that, for example $E(p) = 50$. This corresponds he 0050 sampling plan mentioned earlier (p is approximately 0.11). Because for all Double Zero ıpling plans $c = 0$, the sampling scheme is completely specified once the sample size n is given. each lot size, N, and for each $(n, 0)$ plan, we can calculate (by the OC curve for example) p such $P_a(p) = 0.05$ and for this p compute $E(p)$. Conversely, given N and a particular extent $E(p)$, can find p and hence n such that $P_a(p) = 0.05$. Conveniently, Ellis presents (in his Figure 1) a ɔh which, for various lot sizes, at a given value of $E(p)$, immediately yields the appropriate sample . (Presumably one would round n up to the nearest integer). It is obvious that n decreases in $E(p)$. ndustrial applications 0050 and 0025 plans typically are used corresponding to the tightness of the able-for-use limit.

Note that the presumption of an infinite lot size was necessary in using p to characterize a class ormal distributions as part of the development of the maximum extent condition. Allowing finite ıize means that for some choices of $E(p)$ (hence p) we may not be able to choose n ($\leq N$) such $P_a(p) = 0.05$. This is clearly illustrated in the Ellis Figure. However the more crucial point is the interpretation of this maximum percent of tolerance is fuzzy at best. It is accurate to say that,

given a lot of infinite size, at a quality level p which is acceptable five percent of the time, under normal ity of measurements, the chance of any randomly selected item exceeding $E(p)$ is essentially zero What this has to do with the performance at the same quality level p of *all* items in any particula *finite* lot not necessarily assumed to be from an infinite population which is being continuously samplec is not at all clear. In fact, even if we accept the latter sampling supposition, the distributional assump tions made regarding the measurements will be invalid unless p for the finite lot is assumed to be exactly the same as the p for the infinite lot. This is unreasonable since the value of p will almost surel vary over finite lots, and only for large N would it be approximately the value for an infinite lot. Henc the entire analysis may be very susceptible to error and highly inaccurate when lot sizes are sma (for example the $N = 10, 25, 50$ lines in the Ellis Figure). Thus Double Zero plans only guarantee th indicated protection when lot sizes can be assumed infinite and can be expected to be accurate onl when lot sizes are sufficiently large, i.e., such that the earlier binomial approximation to the hypergeo metric distribution is good.

We also note that the above analysis has been done for any randomly selected measurement fron the assumed independent and identically distributed items in the lot. However, for example, th largest and/or the smallest measurements in the lot will have very different distributions, so the develop ment is not at all valid for the "worst" piece in the lot. Therefore the strength of the second zero in th Double Zero plan may be suspect.

Finally the Double Zero terminology disguises the fact that the analysis has been done for lots of a quality level which we would expect to reject 95 percent of the time. Selection of this 5 percen level of acceptance is arbitrary and is not indicated in the notation. Moreover, the actual quality level, p, at this value of P_a is hidden as well and in fact not clearly defined for finite lot sizes by the above remarks. In any case, quoting the maximum extent at this one quality level hardly gives an indication of the performance of the plan at other quality levels, except for the obvious monotonicity inferences, i.e., for lots of higher quality (smaller p) we will accept more often and the maximum extent will be smaller (with a similar statement for lots of poorer quality).

We now examine the Narrow Limit gaging plans. As we shall see, these plans approach the idea of a "gray" area much more straightforwardly than do the Double Zero plans. Assumptions inherent in the development of such plans are:

(1) We presume that the measured variable follows a normal distribution.

(2) We presume that a lower specification limit, S_L, and an upper specification limit, S_U are given and that these specifications are *wide* in comparison with the process capabilities which under normality can be taken as 6σ.

(3) We presume go—no go gages are prepared which are stricter than specification by an amount $t\sigma$ (t given). Figure 2 illustrates the situation where G_L and G_U denote the gage lower and upper limits respectively. The area between S_L and G_L and between G_U and S_U may be thought of as a "gray" area analogous to the Double Zero plans.

FIGURE 2

Suppose we now consider a process which is producing p percent defectives. Then by our assump-
this area, p, lies completely below S_L or completely above S_U. By symmetry we consider the latter
whereby $S_U = \mu + z_{1-p}\sigma$ and $G_U = \mu + (z_{1-p} - t)\sigma$. Then the chance that an item will not pass the
ow Limit gage is

$$p_t = 1 - \Phi(z_{1-p} - t) = \Phi(z_p + t)$$

e Φ is the standard (unit) normal cumulative distribution function. Any (n, c) sampling plan be-
s a Narrow Limit gaging plan at a given t simply by plotting an OC curve of the probability of
ptance P_a vs. p_t instead of P_a vs. p. In other words, as a result of narrowing the gage by $t\sigma$, we
increased the incoming percent defective from p to p_t. The effect of increasing t is to increase p_t
ence reduce P_a for the lot. These last two sentences express the essence of Narrow Limit gaging
, but, surprisingly, Ott and Mundel fail to directly mention this in their article. Note that if t is
rge so that $t\sigma$ is large relative to p or if p is quite small (virtually zero), then we shall be rejecting
f good quality because the process is producing items outside the Narrow Limit gage but within
pecification limits. The obvious correction for this problem is to either reduce t or increase c.
he Ott-Mundel paper offers little help in selecting for a given t an appropriate n and c except
ggest sizes comparable to those used under a Shewart control plan.

OMPARING THE PLANS

Having described the plans we now compare them. Since all three approaches result in the selec-
of an (n, c) plan, the best comparison among such plans can be accomplished by examining their
curves. This is particularly true in view of the discussion raised in the previous section regarding
heoretical basis for Double Zero sampling plans. In comparing two (n, c) plans, it is clear that at a
value of c the plan with the larger n will always be more stringent (i.e., conservative). Similarly,
fixed value of n the plan with the smaller c will be tighter. Hence, given two plans (n_1, c_1) and
$c_2)$, if $n_1 \geq n_2$ and $c_1 \leq c_2$, clearly the former is superior to the latter. However if $n_1 \geq n_2$, $c_1 \geq c_2$
$\leq n_2$, $c_1 \leq c_2$, the two plans are not immediately comparable. This will frequently be the case
comparing Double Zero plans with Military Standard 105D plans, namely the former will employ
er samples with zero acceptance numbers, while the latter usually will employ larger samples
arger acceptance numbers.

Table 1 illustrates several 0050 and 0025 plans for various lot sizes, acceptance probabilities at
us AQL's, and p such that $P_a(p) = 0.05$. For infinite lot size the 0050 plan calls for a sample of
7 while the 0025 plan calls for a sample size 87. Table 2 offers for these lot sizes several different
ry Standard 105D plans according to a specific AQL along with the acceptance probability at the
and again p such that $P_a(p) = 0.05$. As observed earlier it is not possible for finite lot size to obtain
ble Zero sampling scheme for every percent of tolerance. For example, a lot size of more than
necessary for the existence of a 0025 plan. Comparably for Military Standard 105D procedures,
er preferred AQL's are allowed as the lot size increases. Note that only if the AQL is sufficiently
will the Military Standard 105D plan specify a zero acceptance number and eventually a sample
t least as large as demanded by a Double Zero plan. Hence for many lot sizes, if a moderate AQL
d to determine the Military Standard 105D plan, the Double Zero plan will be superior.

TABLE 1. *Several Double Zero Sampling Plans*

0050 plans						
Lot size	Sample size	Acceptance number	$P_a(1.0)$	$P_a(1.5)$	$P_a(2.5)$	p
50	20	0	0.78	0.68	0.53	10.8
500	23	0	0.79	0.70	0.56	11.5
2000	25	0	0.78	0.69	0.54	11.0

0025 plans						
Lot size	Sample size	Acceptance number	$P_a(.10)$	$P_a(.15)$	$P_a(.25)$	p
150	67	0	0.85	0.79	0.70	3.8
500	78	0	0.92	0.88	0.78	3.5
2000	85	0	0.92	0.88	0.78	3.0

TABLE 2. *Several Military Standard 105D Plans*

Lot size	AQL	Sample size	Acceptance number	P_a(AQL)	p
50	0.65	20	0	0.89	13.9
	1.00	13	0	0.88	20.6
	1.50	8	0	0.88	31.2
	2.50	5	0	0.86	45.1
150	0.10	125	0	0.89	2.4
	0.15	80	0	0.88	3.7
	0.25	50	0	0.89	5.8
500	0.10	125	0	0.89	2.4
	0.15	80	0	0.88	3.7
	0.25	50	0	0.89	5.8
	1.00	50	1	0.90	9.1
	1.50	50	2	0.96	12.1
	2.50	50	3	0.97	14.8
2000	0.10	125	0	0.89	2.4
	0.15	80	0	0.88	3.7
	0.25	200	1	0.91	2.4
	1.00	125	3	0.95	6.2
	1.50	125	5	0.98	8.4
	2.50	125	7	0.98	10.5

Table 3 describes several Narrow Limit gaging sampling plans. In part (a) of the table for various p's (AQL's) the corresponding p_t's are obtained for $t=0.5$, 1,0, and 1.7. These are the t values suggested in the Ott-Mundel article. As can be seen, the effect on the incoming percent defective due to

gaging is rather drastic. In part (b) of the table we consider each of the Military Standard 105D s in Table 2 and notice the effect on P_a at the AQL as the gage changes through $t=0$ (no gaging), 1.0, and 1.7. Again the effect is profound. Clearly, 1.7σ gages are absurdly tight while gages less 0.5σ are hardly worth constructing. Hence the basic fault with Narrow Limit gaging plans is irent; even a modest amount of gaging ($t=0.5$, 1.0) will result in a much too severe sampling plan. eover the situation is seen to grow worse with increasing lot size. The only recourse is to increase ich merely indicates a willingness to accept lots of poorer quality in the first place. At best then, ow Limit gaging when applied to Military Standard 105D plans may be thought of as tightened ection. As a result we devote the remainder of the article to comparisons between the latter plans Double Zero plans.

In particular let us focus on the areas mentioned in the introductory section.

TABLE 3. *Several Narrow Limit Gaging Sampling Plans*

(a)

p	$.5\sigma$ p_t	1.0σ p_t	1.7σ p_t
0.10	0.48	1.8	8.2
0.15	0.68	2.4	10.2
0.25	1.00	3.5	13.4
0.65	2.40	6.9	21.8
1.00	3.40	9.3	26.8
1.50	4.70	12.1	31.9
2.50	7.20	16.9	39.7

(b)

Lot size	AQL	Sampling plan	0σ P_a(AQL)	$.5\sigma$ P_a(AQL)	1.0σ P_a(AQL)	1.7σ P_a(AQL)
50	0.65	(20,0)	0.89	0.68	0.24	0
	1.00	(13,0)	0.88	0.56	0.26	0.02
	1.50	(8,0)	0.88	0.70	0.42	0.04
	2.50	(5,0)	0.86	0.65	0.27	0.09
150	0.10	(125,0)	0.89	0.52	0.11	0
	0.15	(80,0)	0.88	0.63	0.16	0
	0.25	(50,0)	0.89	0.61	0.17	0
500	0.10	(125,0)	0.89	0.52	0.11	0
	0.15	(80,0)	0.88	0.63	0.16	0
	0.25	(50,0)	0.89	0.61	0.17	0
	1.00	(50,1)	0.90	0.48	0.05	0
	1.50	)50,2)	0.96	0.59	0.05	0
	2.50	(50,3)	0.97	0.51	0.02	0
200	0.10	(125,0)	0.89	0.52	0.11	0
	0.15	(80,0)	0.88	0.63	0.16	0
	0.25	(200,1)	0.91	0.39	0.01	0
	1.00	(125,1)	0.95	0.38	0	0
	1.50	(125,3)	0.98	0.40	0	0
	2.50	(125,7)	0.98	0.31	0	0

(1) Regarding underlying assumptions for suitable application, Military Standard 105D plans a applicable to any items inspectable on a dichotomous basis. In the same sense, Double Ze sampling procedures are also widely applicable when considered as ordinary (n, c) plan However only under the further assumptions of normality of measurements and large lot siz will they afford approximately the additionally indicated protection.

(2) Regarding acceptable quality of lots, Double Zero plans strive for zero defectives and hence lot is unacceptable when it contains nonconforming items. The use of zero acceptance numbe insures rejection in the case of known nonconformances. By comparison, Military Standar 105D plans do specify an AQL thereby implicitly allowing acceptance of lots containing know or expected defectives. But these remarks obscure the issue. For any (n, c) plan we ca compute the probability of acceptance at a particularly AQL and it will be greater than zer If one was primarily interested in protection to the level of zero defectives in the lot, pre sumably one would use as small an AQL as possible in selecting a Military Standard 105D pla (and likely obtain a zero acceptance number). Such a choice would essentially negate any ad vantages due the Double Zero plans and in fact might yield a tighter plan (for example a AQL 0.10 in Table 2).

(3) Regarding consumer versus producer protection, the Double Zero plans are designed such tha at a fixed percent of tolerance the OC curves are quite similar regardless of lot size. In thi sense, independent of lot size the consumer is furnished with consistent probability of rejectio of lots with nonconformances. As long as the vendor produces all items to requirements, a lots will be accepted. On the other hand Military Standard 105D procedures, again by specify ing an AQL, furnish the vendor with the knowledge that lots with an acceptable rate of non conformances will be consistently accepted. In selecting a Military Standard 105D plan at an fixed AQL, lot size has a considerable effect on the OC curve (as may be seen somewhat i Table 2 at AQL 1.0). Hence the producer can substantially affect the protection offered by such plans by determining convenient lot sizes. Therefore, proponents of Double Zero plans clain to furnish the consumer with consistent risk while Military Standard 105D plans furnish th producer with consistent risk. But this is an oversimplification. At a given lot size, any samplin plan which is selected by the consumer and then employed regularly, will offer, in terms of its OC curve, consistent probability of acceptance of lots with nonconformances. This protectio can not be abused to the producer's advantage.

(4) Regarding relative cost, the expense for sampling under Double Zero plans will usually be less than for Military Standard 105D plans, since with a zero acceptance number smaller sample sizes will be required to achieve satisfactory protection. In sampling situations where: (i) The vast majority of lots are free of nonconformances; (ii) only a small percentage of lots are non-conforming within the usual range of AQL's; and (iii) the measured variables almost always follow a normal distribution; such sample sizes are more than adequate to detect nonconform-ance in the lot. These remarks support the selection of an appropriately tight (n, c) plan re-gardless of any notions of Double Zero sampling. The cost to the producer in terms of rejected lots is again determined solely by the selection of such an (n, c) plan.

In conclusion, the major advantage offered by Double Zero sampling plans is that, at a given lot size, they tend to be tighter than "usually" employed inspection schemes. However, selection of a Military Standard 105D plan with sufficiently small AQL will in all ways achieve comparable perform-

Furthermore, if tightened inspection or somewhat equivalently Narrow Limit gaging modifica- of Military Standard 105D plans are used, even greater stringency will be achieved.

REFERENCES

Brownlee, K. A., *Statistical Theory and Methodology In Science and Engineering* (John Wiley & Sons, New York, 1965), 2nd ed.

Burr, I. W., *Engineering Statistics and Quality Control* (McGraw Hill, New York, 1953.)

Dodge, H. F. and H. G. Romig, *Sampling Inspection Tables* (John Wiley & Sons, New York, 1959), 2nd ed.

Ellis, E. W., "Double Zero Attribute Sampling Plans." Annual Technical Conference Transactions, American Society for Quality Control (1966), pp. 340–347.

Evans, D. H., "A Statistical Tolerancing Formulation," J. of Quality Technology 2, 226–231 (1970).

Grant, E. L., *Statistical Quality Control* (McGraw Hill, New York, 1972), 4th ed.

Johnson, N. L. and F. C. Leone, *Statistics and Experimental Design: In Engineering and the Physical Sciences* (John Wiley & Sons, New York, 1964).

Military Standard Sampling Procedures and Tables for Inspection by Attributes, MIL–STD–105D, U.S. Government Printing Office, Washington, D.C. (Apr. 1963).

Ott, E. R. and A. B. Mundel, "Narrow Limit Gaging," Industrial Quality Control *X*, 21–8 (1954).

Shewart, W. A., *Economic Control of Quality of Manufactured Product* (D. Van Nostrand Co., Princeton, N. J., 1931).

SUBOPTIMAL DECISION RULE FOR ATTACKING TARGETS OF OPPORTUNITY

Takasi Kisi
Defense Academy
Yokosuka, Japan

ABSTRACT

A player having only a definite number of weapons is hunting targets. His total hunting time is also limited. Targets of opportunity with various values arrive at random, and as soon as a target arrives the player observes the target value and decides whether or not to shoot it down. The issue is what the decision rule is which guarantees him amaximum expected gain during the hunting time. Poisson arrival of the targets, uniform distribution of the target value, and the shoot-look-shoot scheme qre assumed. A decision rule is derived which is not optimal but has a very simple form and gives almost as good value as the optimal decision rule does.

TRODUCTION

Recently D. V. Mastran and C. J. Thomas [7] investigated a decision problem for attacking targets opportunity: A player having only a definite number of weapons is hunting targets. His total hunting ıe is also limited. Targets of opportunity with various values arrive at random, and as soon as a target ·ives the player observes the target value and decides whether or not to shoot it down. The issue is ıat the decision rule is which guarantees him a maximum expected gain during the hunting time. ıstran and Thomas derived relationships of dynamic programming type to obtain the optimal decision e, and solved several interesting example problems.

More or less similar problems were also dealt with by several authors [1–6, 8]. Decision problems this type are formulated rather easily by using dynamic programming, but often it is very difficult to ·ive their closed-form solution. In some cases the numerical solution would be valuable, but the per- ·ective upon the problem would be rather difficult to have with only numerical solutions. Closed-form utions would be of much value for this purpose even though they were not exact solutions but only ɔroximate ones.

In the following, a decision problem is dealt with assuming Poisson arrival of the targets, uniform tribution of the target value, and the shoot-look-shoot scheme of firing. A decision rule is derived ich is not optimal but has a very simple form and gives almost as good value as the optimal decision e does.

SUMPTIONS AND FORMULATION

The problem dealt with below will be based on several assumptions, which we state first.

(1) Targets arrive randomly one by one. The interarrival times between successive targets are umed to be independently and identically distributed with a density function

$$a(t)=\begin{cases}\lambda e^{-\lambda t} & t>0\\ 0 & \text{otherwise.}\end{cases}$$

(2) The value of a target is unknown to the player in advance. He knows only its probability distribution. The value is independent of both the values of the preceding targets and the interarrival time of the target. For simplicity a uniform distribution is assumed here:

$$g(v)=\begin{cases}1 & 0\leqq v\leqq 1\\ 0 & \text{otherwise.}\end{cases} \tag{2}$$

(3) When a target arrives, the player observes its value exactly, and decides whether or not to attack. The shoot-look-shoot scheme is adopted: If the target is worth hunting, he fires a missle with kill probability p, partial damage being ignored. If he secures a hit, another shot is, of course, unnecessary. If he fails, the player is forced to decide whether or not to fire a second time, and a similar procedure, mentioned above, follows. It is assumed that the repeated firings waste no time. Any recall of targets is assumed to be impossible: Once the player decides not to attack a target, any chance for attacking the same target is lost forever.

(4) The player starts his hunting of duration t with a definite number m of missiles. The score of the player is the total value of the targets he will hunt during t. In the following, such a decision rule that maximizes his expected score is named optimal, and this optimal decision rule is sought.

Let $f(t, m)$ be the optimal value function, i.e., the expected score to the player when time t and m number of missiles are remaining and the optimal decision rule is employed throughout the time t. Let $\phi(t, m|v)$ be the similar optimal expected score, on condition that a target of value v has just arrived at t, and let $\Phi(t, m)$ be the optimal expected score, given that a target has arrived at t without any information on its value.

Target value distribution is assumed uniform, and therefore

$$\Phi(t, m)=\int_0^1 \phi(t, m|v)dv \tag{3}$$

and the relationship between f and Φ is obviously

$$f(t, m)=\int_0^t \lambda e^{-\lambda u}\Phi(t-u, m)du \tag{4}$$

or

$$\Phi(t, m)=\frac{1}{\lambda}\frac{d}{dt}f(t, m)+f(t, m). \tag{5}$$

When a target arrives, the player observes the target and finds out its value. Then he makes a decision. The decision rule would take the form "attack the target if, and only if, its value v is not less than a certain value c," where c is a critical level of target value that will be determined later optimally.

First let us consider the case $v\geqq c$. The expected score to the player evaluated at the instant just before his decision is $\phi(t, m|v)$. When the decision is reached a missile is launched, and it hits the target with probability p. As mentioned above, the process runs so quickly that practically no time is

ted. The player gets the value v, and thereafter he is expected to gain the total score $f(t, m-1)$ e time t and $m-1$ of missiles are remaining yet. If the missile misses the target with probability p), the situation changes over to one similar to that just before the decision, but the number of siles is diminished by one. The expected score evaluated at this time point is given by $\phi(t, m-1|v)$. the other hand, the target should be neglected when $v < c$. In this case neither time nor the missile asted, and the expected score at the instant just after the decision is given by $f(t, m)$. Thus,

$$\phi(t, m|v) = \begin{cases} p\{v+f(t, m-1)\} + (1-p)\phi(t, m-1|v) & v \geqq c \\ f(t, m) & v < c. \end{cases}$$

The optimal level c is dependent on both t and m, $c=c(t, m)$, and it is intuitively clear that the ction is increasing in t and decreasing in m. In the following, this property will be utilized without proof.

To see how the player fires and stops, let us suppose the case $v \geqq c(t, m)$. The player dares to the first shot. Suppose he misses the target. The player is on the second decision point where the nber of missiles is diminished by one. As mentioned above, $c(t, m-1) > c(t, m)$, and so if the value arge enough and $v \geqq c(t, m-1)$, he should make the second shot. If $c(t, m-1) > v \geqq c(t, m)$, vever, he should stop.

According to our decision rule, the player having m missiles should attack a target at t if its value ist $c(t, m)$. But because of the continuity property of $\phi(t, m|v)$ in v, the same expected value will scored when the alternative choice, not to attack this target, is adopted here, and the optimal ision rule is employed during the remaining period of time: From (6) we obtain

$$f(t, m) = p\{c(t, m) + f(t, m-1)\} + (1-p)\phi(t, m-1|c(t, m))$$

$$\phi(t, m-1|c(t, m)) = f(t, m-1)$$

h lead to the relationship satisfied by the optimal $c(t, m)$,

$$c(t, m) = \frac{1}{p}\{f(t, m) - f(t, m-1)\},$$

> 0 and $m=1, 2, \ldots$, where $f(t, 0)$ is defined as 0.

Before we solve the equations of the optimal value function, let us mention a word on the asympto- rm of $c(t, m)$.

If the player has a lot of missiles and his hunting time is long enough, or more specifically, if $n \gg 1$, the number of targets which are expected to arrive during the time t is about λt. The number issiles he possesses is m, with which the player can hunt only mp targets because he fires $1/p$ ds on the average to shoot down a target. The player is therefore able to hunt only a fraction t of targets in his total time, and he should select targets with considerably higher value which

belong to the fraction $mp/\lambda t$ from the highest in the target value distribution. Let the critical value be c. Then $1-c=mp/\lambda t$, and so $c(t, m)$ would have an asymptotic form

$$c(t, m) \sim 1-\frac{mp}{\lambda t}. \tag{8}$$

OPTIMAL VALUE FUNCTION

In order to determine the optimal decision rule $\{c(t, m)\}$ the equations satisfied by the optimal value function $f(t, m)$ should be derived and be solved.

From (3) and (6)

$$\begin{aligned}\Phi(t, m) &= \int_0^{c(t,m)} \phi(t, m|v)dv + \int_{c(t,m)}^1 \phi(t, m|v)dv \\ &= c(t, m)f(t, m) + \frac{p}{2}\{1-c(t, m)^2\} \\ &\quad + p\{1-c(t, m)\}f(t, m-1) + (1-p)\int_{c(t,m)}^1 \phi(t, m-1|v)dv\end{aligned}$$

and

$$\begin{aligned}\Phi(t, m-1) &= \int_0^{c(t,m)} \phi(t, m-1|v)dv + \int_{c(t,m)}^1 \phi(t, m-1|v)dv \\ &= c(t, m)f(t, m-1) + \int_{c(t,m)}^1 \phi(t, m-1|v)dv\end{aligned}$$

are obtained, and substituting these relations into (5) yields the following equations.

$$\begin{aligned}&\frac{1}{\lambda}\frac{d}{dt}\{f(t, m)-(1-p)f(t, m-1)\} \\ &\quad = \frac{p}{2}\{1-c(t, m)^2\} - \{1-c(t, m)\}\{f(t, m)-f(t, m-1)\} \qquad m=1, 2, \ldots\end{aligned} \tag{9}$$

It is noted that the condition (7) for the optimal decision has not yet been employed, and (9) is therefore the equation of the expected score $f(t, m)$ corresponding to an arbitrary decision rule $\{c(t, m)\}$.

The equations satisfied by the optimal value functions are from (9) and (7)

$$\frac{1}{\lambda}\frac{d}{dt}\{f(t, m)-(1-p)f(t, m-1)\} = \frac{1}{2p}\{f(t, m)-f(t, m-1)-p\}^2, \qquad m=2, 3, \ldots, \tag{10}$$

and

$$\frac{1}{\lambda}\frac{d}{dt}f(t, 1) = \frac{1}{2p}\{f(t, 1)-p\}^2 \tag{11}$$

for $m=1$. Initial conditions are

$$f(0, m)=0, \qquad m=1, 2, \ldots$$

The solution of (11) with (12) is easily found to be

$$f(t, 1)=p\left(1-\frac{2}{\lambda t+2}\right),$$

analytical solution of (10) seems rather difficult. Then approximate solutions of the following type sought:

$$f(t, m)=p\left\{m-\frac{f_m}{\lambda(t+t_0)}\right\}, \qquad m=1, 2, \ldots$$

ıen (14) is inserted into (10), we obtain a recurrence relation

$$f_m=f_{m-1}+1+\sqrt{2pf_{m-1}+1}, \qquad m=1, 2, \ldots$$

$$f_0=0, \qquad t_0=2/\lambda,$$

n the solution (14) is identical with (13) for $m=1$. For $m \geqq 2$, however, (14) renders approximate ıtions, initial conditions (12) being violated.

Approximate value functions (14) are compared with the exact ones in Figure 1 for some values of where the exact value functions are calculated numerically by use of (10) and (11). For smaller ıe of t the approximate formulae give a bit smaller value, but satisfactory agreement is observed for e t.

BOPTIMAL DECISION RULE

Equation (7) gives the true optimal decision level if $f(t, m)$ is the optimal value function. When roximately optimal $f(t, m)$ is used in place of the true optimal, it will render a suboptimal decision $\{c'(t, m)\}$. Let us adopt the approximate formulae given by (14). Then

$$c'(t, m)=\begin{cases} 1-\dfrac{f_m-f_{m-1}}{\lambda t+2} & t>t_m \\ 0 & \text{otherwise} \end{cases}$$

$$t_m=\frac{1}{\lambda}\left(\sqrt{2pf_{m-1}+1}-1\right)$$

$m=1, 2, \ldots$. Suboptimal $\{c'(t, m)\}$ is compared with $\{c(t, m)\}$ in Figure 2. The suboptimal sion level $c'(t, 1)$ is identical with the optimal, but for $m \geqq 2$ the suboptimal decision levels are

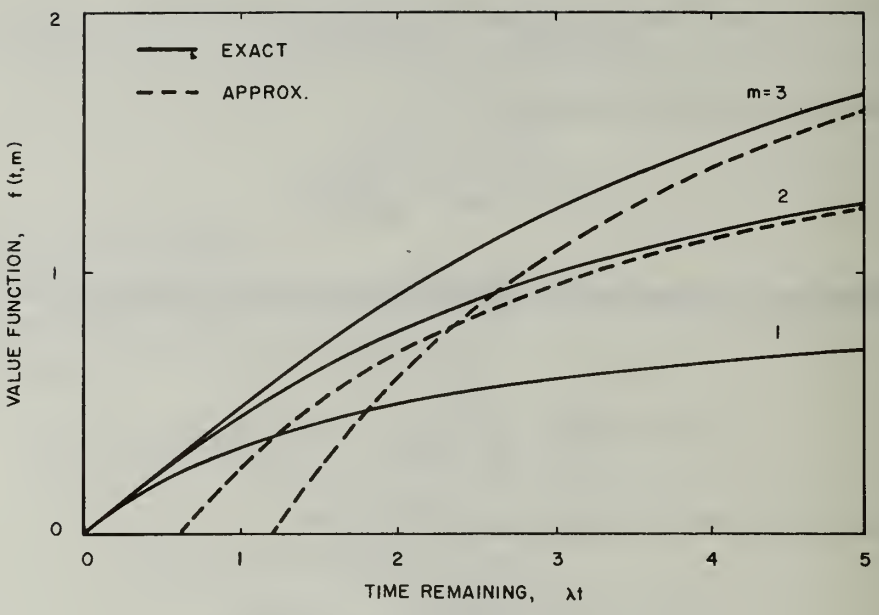


FIGURE 1. Optimal value functions ($p=1$)

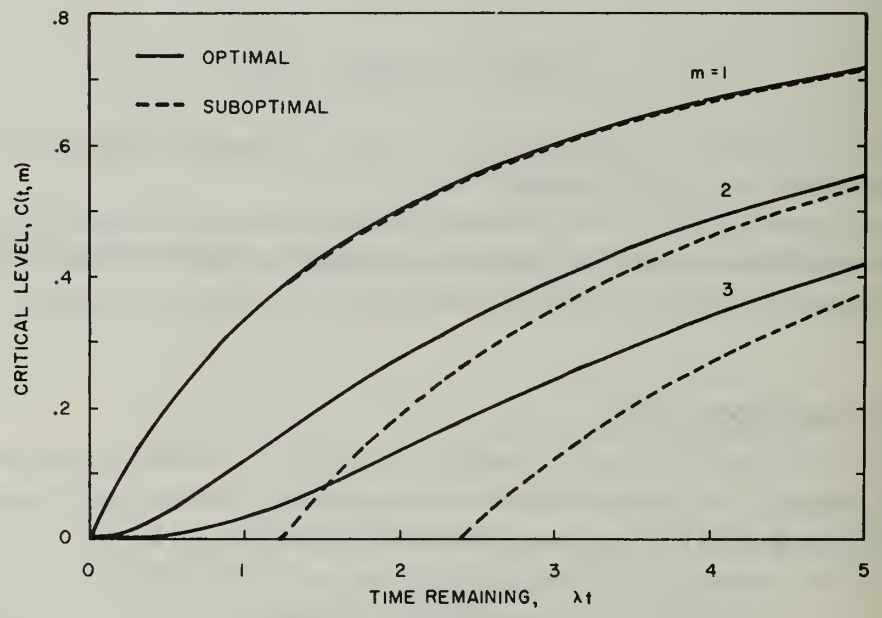


FIGURE 2. Decision rules, optimal vs. suboptimal ($p=1$)

always lower than the corresponding optimal levels as is seen in the figure. The difference is sufficiently small for large t, but is by no means negligible for small t. Nevertheless, the suboptimal decision rule gives very satisfactory expected score even for small t. The solid line curves in Figure 3 are the optimal value functions, and dots are the values when the suboptimal decision rule $\{c'(t, m)\}$ is employed, i.e. f-values calculated from (9), (17) and (15). The difference is negligible. In Table 1 the relative loss in the

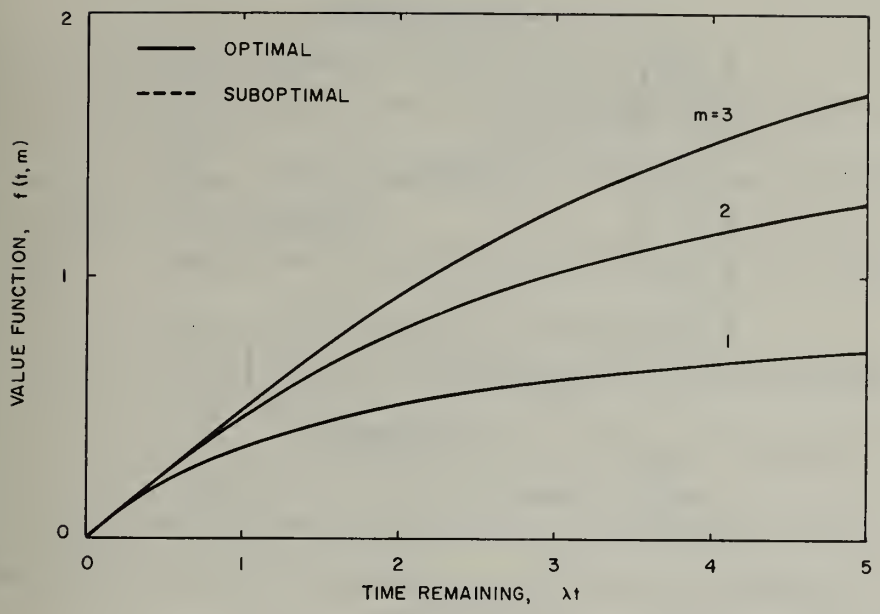


FIGURE 3. Value functions, optimal vs. suboptimal ($p=1$)

ected score of using $\{c'(t, m)\}$ instead of the true optimal $\{c(t, m)\}$ is shown for the case $p=1$. The s is maximum around $\lambda t \sim mp$, but even in this case the value is only 1.23 percent. When t is large $\lambda t \gg mp$, the loss is surprisingly small.

For p smaller than unity the loss grows a little. But calculation shows it never exceeds 3.2 percent. ure 4 tells the fact.

TABLE 1. *Relative Loss in Using Suboptimal Decision Rule in Place of the Optimal.* $p=1$

$\{f_{opt}(t, m)-f_{sub}(t, m)\}/f_{opt}(t, m)\times 100$										
m \ λt	1	2	3	4	5	6	7	8	9	10
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.27	0.71	0.42	0.25	0.16	0.11	0.07	0.05	0.04	0.03
3	0.04	0.60	1.02	0.72	0.47	0.31	0.21	0.15	0.11	0.09
4	0.00	0.19	0.76	1.16	0.91	0.63	0.44	0.32	0.23	0.17
5	0.00	0.04	0.32	0.85	1.21	1.02	0.76	0.55	0.41	0.30
6	0.00	0.01	0.10	0.42	0.90	1.23	1.09	0.84	0.64	0.48
7	0.00	0.00	0.03	0.16	0.49	0.93	1.23	1.13	0.91	0.70
8	0.00	0.00	0.01	0.05	0.22	0.54	0.94	1.21	1.14	0.95

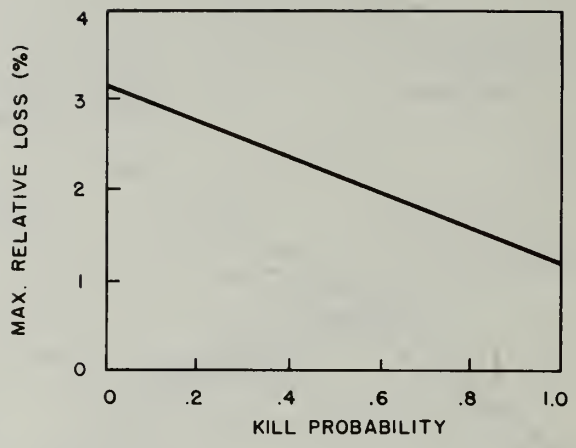


FIGURE 4. Maximum relative loss as a function of p

Finally it is worthwhile to mention that as was expected (17) has the asymptotic form (8) for large t and m. If we define $\Delta_m = f_m - f_{m-1} - mp$, then for $t > t_m$

$$c'(t, m) = 1 - \frac{mp + \Delta_m}{\lambda t + 2}. \tag{19}$$

We have an extra term $2 (= \lambda t_0)$ in the denominator, and an additional term Δ_m in the numerator. The latter is found always positive and $\Delta_m / mp \to 0$ when $mp \to \infty$ as seen in Figure 5.

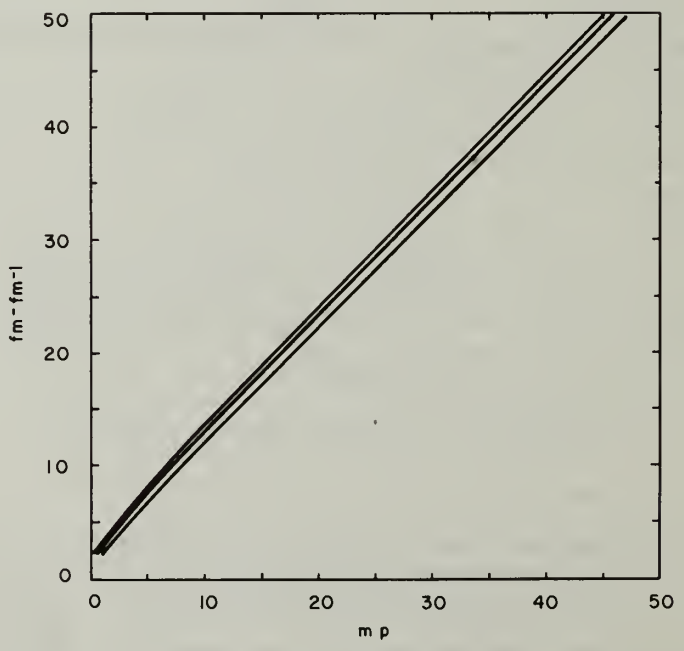


FIGURE 5. $f_m - f_{m-1}$ vs. mp (curves correspond to $p = 1$, 1/2, and 1/4 from the right)

REFERENCES

eGroot, M. H., *Optimal Statistical Decisions* (McGraw-Hill Inc., 1970).

erman, C., G. J. Lieberman, and S. M. Ross, "A Sequential Stochastic Assignment Problem," Man. Sci. *18*, 349–355 (1972).

onis, J. N. and S. M. Pollock, "Allocation of Resources to Randomly Occurring Opportunities," Nav. Res. Log. Quart. *14*, 513–527 (1967).

ilbert, J. P. and F. Mosteller, "Recognizing the Maximum of a Sequence," J. Am. Statist. Assoc. *61*, 35–73 (1966).

oward, R. A., "Dynamic Programming," Man. Sci. *12*, 317–348 (1966).

ōyama, S., "Resource Allocation to Targets of Opportunity," (in Japanese) Master's Thesis, Defense Academy, Japan, Feb. 1972.

astran, D. V. and C. J. Thomas, "Decision Rule for Attacking Targets of Opportunity," Nav. Res. Log. Quart. *20*, 661–672 (1973).

ıkaguchi, M., "A Sequential Assignment Problem for Randomly Arriving Jobs," Rep. Statist. Appl. Res. JUSE *19*, 99–109 (1972).

OCCUPATIONAL STRUCTURE IN THE MILITARY AND CIVILIAN SECTORS OF THE ECONOMY*

Sheldon E. Haber

The George Washington University,

Washington, D.C.

ABSTRACT

This paper focuses on trends in the occupational structure of the military and civilian sectors of the economy. Some implications of these trends for manpower policies for the all-volunteer military establishment are examined.

RODUCTION

The need for a useful and manageable classification of occupations for military manpower manage- is generally recognized. This need has been heightened in recent years as military technology has nced increasing the competition between the military and civilian sectors for skilled manpower. advent of an all-volunteer military establishment further crystalizes the need for occupational ana- wherein the two sectors are viewed as a whole and treated in an integrated manner.

Although there are a number of systems for grouping occupations in the civilian and military ors, the criteria employed in each system are different. The most important system in the civilian or is that developed by the Bureau of the Census. In the military sector, the Department of Defense D) has established a classification scheme which ties together the systems utilized by the individual ary services. The major objectives of this paper are, first, to bring together data from these two ces in order to compare the occupational structure of the civilian and military work force and, nd, to examine the implications of this structure for manpower policies aimed at implementing the olunteer force concept.

While the military comprises only a small proportion of all workers, a substantial percentage of ers in the craftsmen and related occupations are in the armed forces. Additionally, these occupa- are among the most rapidly growing ones in the civilian sector. These findings suggest that policies ted toward enlarging the supply of manpower to the military by increasing lateral transfer from ivilian to the military sector may be constrained by the technology of production in the military or and the supply of skilled manpower in the civilian sector. More appropriate policies would focus aking the military more attractive as a source of craftsmen training and of changing the military uction function in a manner which reduces the requirement for skilled labor.

NDS IN THE OCCUPATIONAL STRUCTURE OF THE CIVILIAN AND MILITARY RK FORCE

One of the more significant changes in the structure of the economy over many decades has been ecline in the proportion of the work force performing manual tasks. Similarly, in the military sector,

This paper was prepared under the Navy Manpower R&D Program of the Office of Naval Research under Contract Number -67-A-0214, Task 0016, Project NR 347-024.

the proportion of personnel attached to ground combat and general duty military occupations has al declined. At the end of World War I, for example, 40 percent of the military work force was estimat to be in this category. [13, p. 52]. By the end of World War II, the corresponding figure was appro mately 24 percent. This may be compared with the percentage of the male workers classified as op atives and laborers (including those in farming) which stood at 33 percent in 1950. Changes in t occupational distribution of the civilian and military sectors since 1940 are shown in Tables 1 and In each table, occupations are aggregated into major groupings based on the Census and DOD clas fication systems, respectively.*

As can be seen from Table 2, during the post-World War II period, the occupational distributi of the military sector has changed considerably. The ground combat and services occupations ha diminished in importance while the electronic and mechanics and repairmen occupations have gain in importance. Of some interest, the proportion of enlisted personnel in ground combat positions d not rise during the Vietnam War, suggesting that this proportion may have fallen as a result of rece reductions in military strength.

Although it is clear that since World War II similar changes have occurred in the occupation structure of the civilian and military sectors, a number of difficulties in comparing the Census and DO data should be mentioned. An obvious difficulty is that there is little overlap in the titles of the maj occupation groupings. The absence of farmers and farm managers from Table 2, for example, is hard surprising. But one notices that professional and managerial workers are also missing from this tabl Some workers who are classified as professionals by the Census, such as photographers and a countants, are classified as "other technical" and "administrative and clerical workers," respectivel

TABLE 1. *Occupational Distribution of Employed Males, 1940–1970*

Occupation group	Percent distribution			
	1940	1950	1960	1970
Professional, technical, and kindred workers	6.1	7.3	9.9	13.5
Managers and administrators, except farm	9.6	10.7	11.0	10.6
Farmers and farm managers	14.8	10.3	5.5	2.7
Clerical and kindred workers	6.0	6.4	6.7	7.2
Sales workers	6.7	6.4	6.9	6.8
Craftsmen, foremen, and kindred workers	14.9	18.6	19.9	19.7
Operatives and kindred workers	17.9	20.0	18.8	18.2
Laborers, except farm	8.9	8.2	7.2	6.1
Farm laborers and farm foremen	8.2	4.9	2.8	1.6
Service and private household workers	6.1	6.1	6.5	7.7
Occupation not reported	0.7	1.1	4.6	5.9
Total [a]	100.0	100.0	100.0	100.0

[a] Rounded to 100.0 percent.

Sources: Bureau of the Census, *Census of Population, 1970, General Social and Economic Characteristics*, U.S. Summary, PC(1)–C1, Washington, D.C., U.S. Government Printing Office, 1972 and *Census of Population, 1950, Characteristics of the Population*, U.S. Summary, Vol. II, Pt. 1, Washington, D.C., U.S. Government Printing Office, 1953.

*The date for the military in Table 2 refer to authorized enlisted personnel strength rather than the actual personnel in ea occupational group. As indicated in [13, p. 193], differences between authorized and actual strength have been small.

TABLE 2. *Occupational Distribution of Enlisted Positions, Selected Years 1945–1967*

Occupation group	Percent distribution			
	1945	1953	1963	1967
Ground Combat	24.1	17.3	14.1	14.1
Electronics	5.8	9.5	14.2	14.7
Other Technical	7.2	7.3	8.1	7.7
Administrative and clerical	15.3	20.6	19.9	18.4
Mechanics and Repairmen	20.0	22.3	24.5	26.1
Craftsmen	9.2	6.6	7.2	6.8
Services	16.6	15.4	11.9	12.0
Miscellaneous	1.9	.....	.....	.....
Total [a]	100.0	100.0	100.0	100.0

[a] Rounded to 100.0 percent.

Sources: Harold Wool, *The Military Specialist: Skilled Manpower for the Armed Forces*, Baltimore, The Johns Hopkins Press, 1969, and Bureau of the Census, *Statistical Abstract of the United States*, Washington, D.C., U.S. Government Printing Office, 1970.

e DOD. More importantly, Table 2 is restricted to enlisted positions and hence excludes officers.* approximately 15 percent of all military personnel are officers [6, p. 25], comparison of the oc- ional distribution of enlisted personnel and civilian workers omits a not insubstantial segment of ork force in the military sector.

Other groups missing from Table 2 are the sales, operative and laborer occupations. Although military personnel are employed in sales occupations, e.g., commissary workers, the counterpart e civilian sales worker is almost absent in the military. Less obvious is the fact that only a small ber of detailed occupations in the military correspond to the operative and laborer occupations. operatives and many laborers are employed in manufacturing, an activity which is almost wholly cted to the civilian sector.¶

The foregoing suggests that in a number of occupation areas there is little direct competition be- the civilian and military sectors. This is not to say that competition is lacking, rather it is indirect t individuals may choose between jobs and careers which are largely unique to the military, e.g., d combat, or largely unique to the civilian sector, notably in laborer, operative and sales occupa- There are a large number of occupations, however, where competition between the civilian and ry sectors is direct, i.e., where occupation skills overlap both sectors. This is true of the adminis- e and clerical as well as the service occupations, but is of particular importance for the craftsmen elated occupations† where shortfalls in the military sector are most likely to occur.**

The DOD classifies officers into eight occupational groups: General officers and executives, tactical operations officers, ence officers, engineering and maintenance officers, scientists and professionals, medical officers, administrators, and procurement and allied officers [5, pp. 1–5]. The second group includes pilots and aircraft crews and is similar to the combat component for the enlisted force. The next four groups are primarily professional workers. The remaining groups se the managerial class.

However, approximately one-quarter of all operatives in 1970 were drivers of automobiles, buses and trucks, a skill which me importance in the ground combat specialties.

the civilian sector, mechanics and repairmen are classified together with other craftsmen. For the discussion that follows, eful to distinguish mechanics and repairmen from other craftsmen as is done in the DOD classification.

The measurement of manpower shortages and surpluses in military occupations is an important but difficult problem. es to remark here that quantitative estimates of imbalances are sensitive to how specific skills are grouped into detailed tions.

TABLE 3. *Annual Rate of Growth in Craftsmen and Related Occupations: Military Positions, 1953–19(and Male Experienced Civilian Labor Force, 1950–1970*

	Enlisted Positions [a]		Experienced Male Civilian Labor Force [a]		Annual Rate of Growth	
	1953	1967	1950	1970	Enlisted Positions	Exp. Civ. Male Labor Force
Craftsmen and related occupations	872	1,130	8,423	11,503	1.9	1.6
Electronics technicians [b]	216	349	107	307	3.5	5.4
Mechanics and repairmen [c]	506	619	1,646	2,196	1.5	1.5
Aircraft	172	236	73	144	2.3	3.5
Automotive	82	106	667	928	1.9	1.7
Other	252	277	[c] 906	[c] 1,124	0.7	1.1
Craftsmen (excluding electronics technicians and mechanics and repairmen)	150	162	7,037	9,000	0.6	1.2
Construction and utilities [d e]	51	62	3,163	3,526	1.4	0.5
Shipboard operations [f]	36	34	59	30	−0.4	−3.3
Metal working [g]	24	23	1,008	1,103	−0.3	0.5
Other	39	43	2,807	4,341	0.7	2.2
All other occupations [h]	1,394	1,245	34,299	37,909	−0.8	0.5
Total	2,266	2,375	42,722	49,412	0.3	0.7

[a] Number in thousands.

[b] Includes radio and television mechanics.

[c] Excludes air conditioning, heat and refrigeration mechanics, and radio and television mechanics.

[d] Includes stationary firemen, power station operators, construction managers and construction apprentices.

[e] Includes air conditioning and heat and refrigeration mechanics.

[f] Includes sailors, and deckhands and boatmen.

[g] Includes welders and flamecutters, and metalworking apprentices.

[h] Excludes occupations cited in footnotes d, f, and g.

Source: Bureau of the Census, *Census of Population, 1970, Detailed Characteristics*, U.S. Summary, PC(1)–D1, Washington, D.C., U.S. Government Printing Office, 1973 and *Census of Population, 1950, Characteristics of the Population*, U.S. Summary, Vol. II, Pt. 1, Washington, D.C., U.S. Government Printing Office, 1953.

In Table 3, the craftsmen and related occupations are separated into three categories: Electron technicians, mechanics and repairmen, and craftsmen (excluding mechanics and repairmen). Includ among electronics technicians are radio operators who in the civilian sector are classified by the Cens as professional workers. Likewise, included among craftsmen (excluding mechanics and repairme are some workers, principally in the construction industry, who are classified as operatives and manage by the Census.*

A number of interesting inferences can be drawn from this table. The craftsmen and related c cupations now comprise almost one-half of all enlisted positions (and approximately 40 percent of t entire military work force including officers). In the civilian sector, about one-fifth of the experienc male civilian work force† are in these occupations. Hence, although the experienced male civili labor force is more than 20 times as large as the enlisted military work force, the number of civili craftsmen is only 10 times as large as the number of craftsmen in the military. Additionally, more th

*Civilian occupations corresponding to military groupings shown in Table 3 are found in [13].

†The experienced labor force includes the employed and unemployed persons who have worked at any time in the pa

nalf of the craftsmen in the military are concentrated in mechanics and repairmen occupations. In e occupations, the ratio of civilian to military workers is only 3.5 to 1. Thus, for this important p of occupations, labor market conditions in the civilian economy will have important implications nanpower management in the military sector, and, conversely, manpower policies in the latter or can have an important impact on the former sector. For example, in a draft environment, much e training provided by the military will be lost when draftees return to civilian life and embark on ers of their own choice. In an all-volunteer military establishment where individuals are much more y to receive the type of training they desire, there is a correspondingly greater likelihood that if choose a career in the civilian sector, it will be in a field related to their military training.*

The data in Table 3 indicate that for some occupational groups the ratio of civilian male workers litary workers is high, e.g., in construction and utilities and the metal working trades. Where this the possibilities for lateral transfer between the civilian and military sectors may be good. Where atio is low, however, the pool of manpower in the civilian sector available for work in the military ely to be small and lateral transfer may be difficult to effect. It is of some interest, therefore, to that for the electronics and aircraft mechanics and repairmen occupations, which comprise more one-half of the positions in craftsmen and related skills in the military, the ratio of civilian to ıry workers is very low.

Lateral transfer may be difficult to effect for another reason despite its apparent attractiveness to ıilitary. The cost of training to the military can be quite high, as much as $15,000 in some spe-es, but more typically $10,000–$12,000 (see [7]). If this amount is amortized over a 4-year period, 0–$3,000 per year could be offered as a premium to workers trained in the civilian sector who t a temporary job assignment in the military. Lateral transfer is not without its costs to the dual, however. Besides the possibility of some unemployment when the military contract period there is the possibility of foregone promotion opportunities, lost seniority and pension rights, ıe psychic costs of changing jobs, particularly if a change in location is required. These costs can ge. For example, Galloway [8, pp. 35–39] estimates that the cost of changing from a job in New nd to one on the Pacific coast, most of which is unrelated to the expense of moving family posses-ranges from $3,000 to $4,000 per year over a 5-year time horizon. The point to be emphasized s that in evaluating lateral transfer as a means for increasing the supply of manpower to the ry, costs to the individual as well as benefits to the military need to be taken into consideration. ı addition to the limited supply of civilian manpower in the craftsmen and related occupations e to the needs of the military, the growth of these occupations in the civilian sector has been apid. During the period since World War II, the rate of growth of civilian craftsmen and related ations has been 1.6 percent per year compared to 0.5 percent per year for the rest of the civilian force.† To illustrate the cumulative effect of this difference in growth rates, $100 compounded lly at a rate of 1.6 percent will grow to $200 in 44 years. Compounded at an annual rate of 0.5 ıt, it would take 139 years to double. It should be noted, moreover, that the most rapidly growing

---

*This suggests that earlier studies of transference of skills [12] and assessments of the extent to which military training fic or general training may not be relevant in the context of a no-draft environment. If transference of skills does occur, al and private returns on investment in human capital, represented by training in the military in craft trades and other as, could be large.

†Annual rates of growth are shown since the data in each sector cover different periods of time. The annual rate of growth uted by solving for r in the relationship $Y=A(1+r)^t$ where A and Y are the number of individuals in an occupation at al and terminal dates, respectively, and t is the number of years between the two dates.

craftsmen occupations in the civilian sector—electronics technicians and aircraft mechanics repairmen—are also the occupations for which the demand for manpower has risen most rapidly in military. Additionally, demand in these occupations is increasing faster in the civilian than in military sectors. Hence, it is not surprising that the military has experienced difficulty in meeting needs for personnel in these technical fields. Granting that the drawdown from the peak strength of Vietnam War will improve the short-run military manpower balance sheet, should these trends continue, they portend serious problems for the all-volunteer force which are now somewhat obscured by the emphasis on procurement of combat arms personnel.

The trends noted raise another important issue for long-run manpower policy, namely, whether the military should attempt to maximize initial enlistments or reenlistments.* At least two arguments can be offered for the latter policy. First, the recent decline in the birth rate will make future recruitment of new entrants into the military sector more difficult (see [11]). Second, a military establishment comprised of a large porportion of career personnel would reduce the cost of training an all-volunteer force. But as indicated above, the demand for personnel with craftsmen skills by the military is large relative to the total demand for such services, and civilian craftsmen and related occupations have been growing more rapidly than other occupations,† particularly in the technical skill areas. Hence, the marginal cost to the military of attracting additional *career* personnel could be substantially higher than average cost, particularly if the mechanism used to obtain personnel is across-the-board pay raises versus selective pay raises or bonuses.

Although much attention has been given to the wage elasticity of enlistment (see [1], [3], [1 there is some indication that the opportunity to learn a skill or trade is at least as important an inducement to enlist as is military pay. For example, in the Gilbert Youth Survey [4, p. 68], skill/ training is cited almost as often as compensation as the incentive, if any, which would induce enlistment. When attention was restricted to youths who indicated they might enlist at some time, train was rated by twice as many respondents as compensation as an incentive which would exert a stro influence on the decision to enlist [4, p. 79]. This suggests the policy of utilizing the military as a vehi for training a large number of young people for employment in areas where shortages are being perienced in the civilian sector (as in the craftsmen occupations) as an alternative to relying on enlistments as the major source of military manpower. This policy could provide a means for offsett the potential decline in manpower availability by increasing the propensity of those who are availa to enlist for military service. It would also reduce the possibility of spiraling wage rates which co result from both sectors competing for the same body of workers to fill career positions. By provid initial training to inexperienced workers rather than attempting to maximize reenlistments, su competition would be reduced and retention costs lowered.

The policy of using the military to provide on-the-job training for young people as a means satisfying the manpower requirements of a volunteer military establishment is similar in some respe to Project 100,000 except that the emphasis is directed not so much at providing training to individu

*In one sense, the two options are compatible in that the larger the number of initial enlistments, the larger the numb of reenlistments, all other things remaining equal. In another sense, they are not. The rapid rise in the cost of military manpow from 42 to 56 percent of the military budget between 1968 and 1974 [2], suggests that in the future it may be necessary to cho between making initial or career service more attractive, but not both. Recent discussions in Congress concerning the burgeon costs of military pensions are one indication that this choice may not be far off.

†The same is also true for professional, technical and kindred workers. During the period 1950–70, the annual rate of crease of male experienced workers in this occupation group was 4.2 percent.

otherwise would have difficulty in competing in the labor market as to individuals who can compete who would normally opt for training in the civilian sector. Some indication that individuals and ain institutions recognize the employment potential of craftsmen work is suggested by the in- sing number of colleges which are offering vocation-oriented training.

The explicit policy of making craftsmen training in the military a more attractive option to young ple would likely increase the defense budget. As an offset, however, this approach could ease the lem of maintaining a trained reserve to meet contingency situations, not to mention the substantial al and economic benefits that might be realized.

NCLUSION

In this paper, the occupational structure of the civilian and military work force is examined, some implications are drawn pertaining to military manpower policies in an all-volunteer environ- t. The occupation data reveal that over time the less skilled components of the military work force e been declining, paralleling changes in the composition of the civilian labor force. While this has n known for some time, it is less well recognized that employment in the craftsman and related s in the military comprises a substantial proportion of total employment in these skills in the economy rge. Moreover, male employment in the craftsmen and related occupations in the civilian sector been growing at a substantially faster rate than total male employment. These trends suggest that civilian and military sectors are competing for a limited supply of male workers, and hence raises rtant questions for manpower policy. Perhaps the most crucial issue is whether the military should its needs for craftsmen and related skills through increased retention or through training programs d at attracting new entrants. Both policies are likely to have a high price tag. The latter policy he virtue of providing a pool of trained, skilled manpower to the civilian sector and the potential eeting military reserve requirements. The occupational data also suggest that although the benefit e military of lateral transfer from the civilian sector is high, the possibilities of lateral transfer may mited, since in those occupations where manpower shortages are most likely to occur, the ratio of an to military employment is very low.

Given the constraint that occupational structure places on policies designed to increase the ly of military manpower, the question arises as to the options that are most likely to succeed in ng the long-run manpower needs of an all-volunteer force. One obvious way to correct labor market lances is to increase compensation in occupations where a shortage of personnel exists. However, may be costly if pay increases are directed toward increasing retention. An alternative approach d focus on first-term enlistees and emphasize the noncompensation inducement of training. The t of this approach is to increase the supply of skilled manpower to the military. Adjustments can ade also on the demand side. When the price of a factor input rises, or a shortage manifests itself se the price of factor is sticky, as is the case in military labor markets where market forces te imperfectly, output levels can be maintained most efficiently by substituting other factors for ven factor. Where the given factor is skilled labor, this generally means substituting capital for to reduce the requirement of labor at all skill levels. Capital substitution has been taking place military increasing output, e.g., firepower, per unit of labor input. The occupational data reviewed s paper suggest, however, that the capital that has been developed has reduced the requirement for lled personnel but at the same time has increased the requirement for more skilled personnel.

Thus, in the future, emphasis needs to be given to changes in technology which lead to the substitut of less skilled labor for more skilled labor and reduces the requirement for the latter in absolute term

Notwithstanding the recent attention that has been focused on the difficulty of meeting grou combat personnel requirements in the Army, this study suggests that in the near future, the servi may have even greater difficulty in meeting requirements for skilled personnel in craftsmen a related occupations.

Acknowledgment

Thanks are due to Tulay Demirler whose assistance in the research was most helpful.

REFERENCES

[1] Altman, Stuart H., "Earnings, Unemployment and the Supply of Enlisted Volunteers," *Journ of Human Resources*, IV–1, 1969, 38–59.

[2] Carter, Luther J. *et al.*, "Fiscal 1974 Budget," *Science*, *179*, 9 February 1973, 550–1.

[3] Cook, Alvin A., Jr., "The Supply of Air Force Volunteers," The RAND Corporation, Santa Moni California, 1970.

[4] Department of Defense, *Attitudes of Youth Toward Military Service: Results of National Surve Conducted in May 1971, November 1971, and June 1972.* Manpower and Reserve Affairs, Rep No. MA 72–7, August 1972.

[5] Department of Defense, *Officer Occupational Conversion Table*, Office of the Assistant Secreta of Defense, Manpower and Reserve Affairs, DA PAM 611–11, March 1972.

[6] Department of Defense, *Selected Manpower Statistics*, OASD (Comptroller), Directorate for I formation Operations, April 15, 1973.

[7] Department of the Navy, *Annual Training Time and Cost for Navy Ratings and NEC's*, Bure of Naval Personnel, NavPers 18660, November 1962.

[8] Gallaway, Lowell E., *Geographic Labor Mobility in the United States, 1957 to 1960.* (Washingto D.C.: Government Printing Office, 1969).

[9] Gillete, Robert, "Military R and D: Hard Lessons of an Electronic War," *Science*, *182*, 9 Novemb 1973, 559–561.

[10] Gray, Burton C., "Supply of First Term Enlistees," in The President's Commission on an A Volunteer Force, *The Report of the Presidential Commission on an All-Volunteer Force*, I 2–1–40. (Washington, D.C.: Government Printing Office, November, 1970).

---

*The development of automated ships is an example of a technological change which can lead to a reduced demand skilled manpower. The Air Force also appears to be giving serious consideration, and perhaps the other services as well, to cap equipment which employs less skilled labor input, e.g., the use of unmanned, expendable drones instead of manned airc (see [9]).

Stewart, Charles T., Jr., "Demographic Trends and Naval Manpower Policies," The Graduate School of Arts and Sciences, The George Washington University, Technical Report Serial TR-1122, 8 June 1973.

Weinstein Paul A., and Eugene L. Jurkowitz, "The Military as a Trainer: A Study of Problems in Measuring Crossover," *Proceedings of the 19th Annual Winter Meeting, Industrial Relations Research Association*, San Francisco, California, December 28–29, 1966.

Wool, Harold, *The Military Specialist: Skilled Manpower for the Armed Forces*, Baltimore, The Johns Hopkins Press, 1969.

ARMOR CONFIGURATIONS VIA DYNAMIC PROGRAMMING

R. E. Shear, A. L. Arbuckle, V. B. Kucher
USA Ballistic Research Laboratories
Aberdeen Proving Ground, Maryland

ABSTRACT

The problem of selecting materials, their thicknesses and order for armor designed for the defeat of shaped charge threats, has been formulated as a constrained optimization problem. The mathematical model provides an optimal order and thickness of each layer of material such that the resulting armor configuration will be of minimum mass per unit area subject to constraints on total thickness and shaped charge jet tip exit velocity.

NTRODUCTION

In the design of vehicular armor, the designer is confronted with a bewildering array of constraints possibilities for the armor configurations. Given a vehicle, the designer must consider the threats fend against, materials to use, and how to construct or fabricate the armor. Constraints are imposed rms of thickness, weight, and technology as well as location of vulnerable areas and components. In the following, it will be assumed that the threat is due to a shaped charge and that the armor is cated out of layers of preselected material of various thicknesses. The primary damage mechanism e shaped charge is the high velocity metallic jet which is formed upon collapse of the metallic al liner in the charge. This jet can penetrate and, actually, perforate several inches of armor. The ining jet, after perforation, may cause fuel and ammunition fires, incapacitate or kill crew members, mage vulnerable components. In addition, spallation of the back surface of the armor usually nces the damage capability of the shaped charge. The reader interested in the mechanism of forma- history, etc., of the shaped charge is directed to the paper of Birkhoff et al. [2] and the references therein.

The selection and order of materials to be used in an armor fabricated of N-layers of material for gainst shaped charge threats will be formulated as an optimization problem. In particular, two mic programming problems have been formulated for the design of layered armor. Both models imple penetration equations for calculating jet tip velocity, thus interface phenomena, jet particula- etc., are not described. Such equations and the resulting models, however, provide a good approxi- n of the exit velocities of continuous jets. The models, as developed, provide an effective rationale e preliminary design stages of such armor.

ORMULATION OF MINIMUM MASS ARMOR PROBLEM

In this section the problem of selecting materials, their thicknesses, and their order such that esulting armor configuration will be of minimum mass per unit area will be formulated. The num will be subject to constraints imposed upon the total armor thickness and on the jet tip ity.

f T denotes the total thickness of the armor; t_i, the thickness of the ith layer; ρ_i, the density of h layer; Z, the standoff distance; V_i, the initial striking velocity of the jet, and V_e, the exit jet tip ity, then we seek to minimize the total mass per unit area of the armor, i.e.,

$$\underset{\{t_i,\,\rho_i\}}{\text{Minimize}} \sum_{j=1}^{N} \rho_j t_j \tag{1}$$

subject to the constraints

$$\sum_{j=1}^{N} t_j = T, \tag{2}$$

$$V_e/V \leqslant A. \tag{3}$$

and

$$\rho_i \epsilon \{\rho^{(1)}, \rho^{(2)}, \ldots, \rho^{(k)}\} \tag{4}$$

where $\rho^{(i)}$ denotes the density of the ith material. The $\rho^{(i)}$'s, or the materials, are preselected, and th ρ_i's and t_i's are to be selected so as to minimize the total mass per unit area subject to the impos conditions. The jet tip, impacting slab i with velocity V_i, is eroded during the penetration process a the residual jet exits, if it perforates slab i, with exit velocity V_{ei}. The exit velocity is given by th empirical relation

$$V_{ei} = V_i g_i(Z_i, t_i, \rho_i).$$

For most metals the function g is of the form:

$$g_i \equiv \left(\frac{Z_i}{Z_i + t_i} \right)^{r_i} \tag{5}$$

where Z_i is the standoff distance to the surface of the ith layer, $r_i = \sqrt{\rho_i/\bar{\rho}}$, ρ_i is the density of th ith slab, and $\bar{\rho}$ is the density of the jet material. Note that, for layered armor, the exit velocity of th jet, after perforating the ith layer, becomes the striking velocity at the $(i+1)$ layer. The constrai equation (3) can then be written as

$$V_e/V = \prod_{i=1}^{N} g_i(Z_i, t_i, \rho_i) \leqslant A \tag{6}$$

The minimization problem described by equations 1, 2, 4, and 6 will be cast as a dynamic progra by noting that the minimum will depend upon A, Z and T. These become the state variables of th dynamic programming formulation. The decisions are to select the t_i's and ρ_i's and the optimal polic is that collection $\{\rho_1^*, t_1^*, \rho_2^*, t_2^*, \ldots, \rho_N^*, t_N^*\}$ which yields the minimum for given values of A, T and Z The formulation as a dynamic program depends upon "The Principle of Optimality" as given b Bellman [1], viz.,

> "An Optimal Policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision."

Let $H_N(T, Z, A)$ = the minimum mass per unit area of an N-layered slab of armor of total thickness th the property that the ratio of the exit velocity of the residual jet to the initial jet striking velocity ss than or equal to A for a shaped charge at a standoff distance Z.

Relating the problem formulation and the definition of H_N to the Principle of Optimality, we have nitial state (T, Z, A), the initial decisions, t_N and ρ_N, when the layer N is taken to be the slab, e surface is at a standoff distance Z, and the resultant state (T', Z', A') resulting from the initial sion, defined by $T'=T-t_N$, $Z'=Z+t_N$, $A'=A/g_N(Z, t_N, \rho_N)$. The remaining decisions, i.e., $t_{N-1}, \ldots, \rho_1, t_1$, must constitute an optimal policy with respect to the state (T', Z', A') where $T-t_N$, $Z'=Z+t_N$ and $A'=A/g_N(Z, t_N, \rho_N)$. That is, the remaining decisions constitute the mizing solution for the problem

$$\underset{\{t_i, \rho_i\}}{\text{Minimize}} \sum_{i=1}^{N-1} \rho_i t_i$$

ect to

$$\sum_{i=1}^{N-1} t_i = T'$$

$$\prod_{i=1}^{N-1} g_i(Z_i, t_i, \rho) \leq A'$$

the ρ's belong to the given set of material properties. The solution of the above problem is given $_{N-1}(T', Z', A')$.

The recursive relationship for the dynamic programming formulation is readily obtained by noting the contribution to the mass per unit area which results from an initial decision $\{\rho_N, t_N\}$ is $\rho_N t_N$. rding to the principle of optimality, the remaining decisions must constitute an optimal policy; fore, the minimum mass per unit area, with respect to the resultant state is $H_{N-1}(T', Z', A')$. , making the best initial choice yields the recursive relationship:

$$H_N(T, Z, A) = \min_{t_N} \min_{\rho_N} \{\rho_N t_N + H_{N-1}(T-t_N, Z', A/g_N(t_N, \rho_N))\}$$

Equation (7) is the recursive relationship of the dynamic program and is the computational basis he solution. The state variables can assume any value of interest, thus resulting in the embedding e original problem into a family of two-dimensional minimization problems.

A METHOD OF COMPUTATION

The dynamic program, given by (7), involves three state variables which, in general, adds to the plexity of the computation, or, as Bellman describes it, the "curse of dimensionality". In this prob- however, the curse is not too damaging in that the special structure of the problem allows con- able reduction in the amount of calculation necessary to solve the problem. For example, if Z T' are given, then one readily sees that throughout all stages of the calculation the functions need

only be computed for values of T' and Z' such that $T'+Z'=T+Z$. Additional time and computation savings may also be achieved by noting the discontinuities of the functions H_N. The functions H_N a piecewise constant and have points of discontinuity due to the discrete choices of the ρ's or the choi of allowable materials. The work of Haymond (3) is important at this stage, for Haymond conside the discontinuity sets of the return function which are of the same form as the H_N functions her The constraint set of equations is different from that of Haymond; however, it is conjectured th Haymond's results can be readily transformed to the problem at hand. In short, if the discontinui set of H_N is denoted by $D(H_N)$, the discontinuity set of g_i by $D(g)$, and if $D(g)*D(H_N)$ is the s of products of the elements of the discontinuity sets of H_N and g_i, then we have $D(H_N) \subset D(H_{N-}$ $\cup D(g) \cup (D(g)*D(H_{N-1}))$.

Thus, by saving the discontinuities of H_{N-1} and g, candidates for points of possible discontinuiti of H_N can be predicted. This normally results in fewer function evaluations and a correspondi reduction in computation which may be substantial if the discontinuities are widely separated and one uses an effective algorithm for computation. In this problem, the discontinuities may be extreme close and numerous depending upon N, the number of layers; therefore, in the present calculation, course grid was used coupled with a modified bisection method to locate the points of discontinuit No comparisons, in terms of speed of computation, have been made between the two methods, but is clear that more function evaluations must be made with the latter approach.

Equation (7) is a recursive formula where the solution at stage N, or layer N, depends upo stage $N-1$. The computation can proceed given the value of $H_1(T, Z, A)$. For simplicity in expositio it will be assumed that the material set (4) consists of metals only; thus the g-functions given by (can be ordered with respect to the densities. Since $0 \leq Z/(Z+t_i) \leq 1$ for all Z and t_i, g_i decreases wit increasing density, i.e., if $\rho_i > \rho_j$, then $g_i < g_j$. The set of material represented by (4) shall be assume to be ordered such that $\rho_1 < \rho_2 < \ . \ . \ . \ < \rho_k$. H_1 can then be defined as

$$H_1(T, Z, A) = \begin{cases} \rho_1 T & \text{if } g_1(T, Z) \leq A \\ \rho_2 T & \text{if } g_2(T, Z) \leq A < g_1(T, Z) \\ \cdot & \\ \cdot & \\ \cdot & \\ \rho_k T & \text{if } g_k(T, Z) \leq A \leq g_{k-1}(T, Z). \end{cases} \tag{8}$$

For values of $A < g_k(T, Z)$, $H_1(T, Z, A)$ is defined to be infinite, indicating that for these value of A, no material exists within the preselected material represented by (4). The values given by (8 are not, in general, applicable if materials other than metals are used; the main point is that $H_1(T, Z, H$ can be calculated, thus permitting the stage-by-stage computation. Later in this report configuration for spaced armor will be computed which will modify the calculation at alternate stages; 2, 4, 6, . . . where the decision will include the possibility of the even numbered layers being air only. The g func tion for air is slightly different from that of (5) but causes no computation difficulty.

Given H_1 (T, Z, A) for selected values of T, Z and A thus permits calculations of H_2 $(T, Z, A$ The calculation then proceeds sequentially with N. For example, if $T=T_0$, $Z=Z_0$, and $A=A_0$, and it is desired that $H_2(T_0, Z_0, A_0)$ be calculated, then we can use (7) and (8). In particular, suppose tha

aterial set consists of two materials (metals) of density $\rho^{(1)}$ and $\rho^{(2)}$, respectively. Furthermore, ne that each layer thickness must be a multiple of $\Delta t = T_0/3$. Thus, t_2 can assume the value of $2 \cdot \Delta t$ since an armor of two layers is required. The initial decision is to choose the material density yer two, $\rho^{(1)}$ or $\rho^{(2)}$, and the thickness $t_2 = \Delta t$ or $2 \cdot \Delta t$ of layer two. For example, if we choose d $\rho^{(1)}$, then the immediate return, i.e., the mass of the 2d layer, is $\rho^{(1)} \cdot \Delta t$, and, since our remaining tion or policy must be optimal, we have, with this choice, the armor mass per unit area,

$$\rho^{(1)} \cdot \Delta t + H_1 \left(T_0 - \Delta t, Z_0 + \Delta t, A_0/g_2 \left(\Delta t, \rho^{(1)}, Z_0\right)\right).$$

or each choice of an initial decision ρ_2, t_2 we obtain an expression or value similar to (9). By izing over all such decisions, $H_2(T_0, Z_0, A_0)$ will be obtained. Examination of this example that, in order to obtain $H_2(T_0, Z_0, A_0)$, values of H_1 must be known at points

$$\left(T_0 - \Delta t, Z_0 + \Delta t, A_0/g_2 \left(\Delta t, \rho^{(1)}, Z_0\right)\right)$$

$$\left(T_0 - \Delta t, Z_0 + \Delta t, A_0/g_2 \left(\Delta t, \rho^{(2)}, Z_0\right)\right)$$

$$\left(T_0 - 2 \cdot \Delta t, Z_0 + 2 \cdot \Delta t, A_0/g_2 \left(2 \cdot \Delta t, \rho^{(1)}, Z_0\right)\right)$$

$$T_0 - 2 \cdot \Delta t, Z_0 + 2 \cdot \Delta t, A_0/g_2 \left(2 \cdot \Delta t, \rho^{(2)}, Z_0\right)\Big).$$

$H_2(t_0, Z_0, A_0)$ is then obtained by selecting the smallest value of return values given by expres- (9) evaluated at the above points. In similar fashion, $H_3, \ldots, H_N$ can be obtained.

ARMOR CONFIGURATIONS OF MINIMUM MASS

he dynamic programming formulation and solution of the minimum mass per unit area problem, ions (1–4), provides information on the effect of total armor thickness, minimum layer thick- standoff distance, number of layers, etc., upon the minimum mass per unit area of the armor. In words, the dynamic program, Equations 7 and 8, includes, in the course of computation, a sen- y analysis. In this section, some of these effects will be illustrated.

low, Equation 7 is solved recursively, first for $N = 1$, then $N = 2$, etc.; therefore, we first look at iriation of minimum mass per unit area as a function of the number of layers, N. This effect is ated in figure 1 for an armor of total thickness, $T = 10.5$ cm, standoff distance, $Z = 4.0$ cm, and ty ratio $V_e/V \leq 0.4$. In this calculation each layer thickness is an integral multiple of 1.5 cm, that minimum layer thickness equals 1.5 cm. The maximum number of layers, therefore, is seven, ach layer is 1.5 cm thick. In addition, for this example the density set (4) has densities corres- ig to those of air, beryllium, aluminum, titanium and iron (steel). An additional constraint is im- upon the solution in that air is considered as a candidate at layer n if, and only if, n is even. For f the armor configurations given in figure 1, the ratio of velocities is less than or equal to 0.4, that armor "defeats" the same threat. The mass/area number denotes the armor of minimum mass it area. For example, for $N = 1$, steel would also "defeat" the given threat but the mass/area be substantially higher than that of titanium. We note from figure 1 that the minimum mass is a reasing function of N.

he minimum mass per unit area is also dependent upon the preselection of materials and in figure llustrate this dependency. The first armor illustrated, again for the same condition as of figure 1, a minimum of $42.4 g/cm^2$ when the material density set consists only of densities of beryllium,

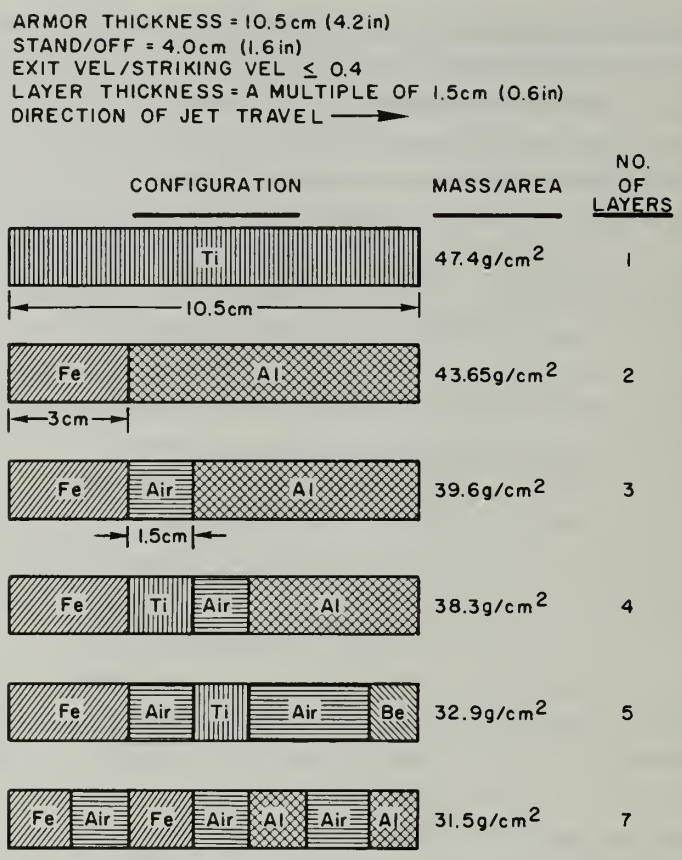


FIGURE 1. *Minimum mass/area vs. number of layers.*

aluminum, titanium and steel. The second consists of steel, aluminum and air and we note the d crease in the value of the minimum. The third armor is for the same class of material as of figure 1 a we note again the decrease in the value of the minimum.

The minimum layer thickness also has an effect upon the values of the minimum. The minimu layer thickness is the value of the increment thickness, a decision value, and results from the di cretization of the thickness. Of course, the minimum layer thickness may also be a technological fabrication constraint as well. The effect of minimum layer thickness is illustrated in Table I for $\Delta t = 0.$ 1.0, and 1.5 cm. In general, we see that the minimum is a nonincreasing function of Δt.

TABLE I.

Effect of Minimum Layer Thickness, ΔT, Upon Minimum Mass/Area (g/cm^2)

$T = 9$ cm ; $Z = 4$ cm

$V_e/V \leq 0.4$

N \ ΔT	0.5 cm	1.0 cm	1.5 cm
1	70.20	70.20	70.20
2	44.70	44.70	47.20
3	40.35	41.70	43.20
4	39.15	39.30	43.20
5	35.25	36.60	39.15

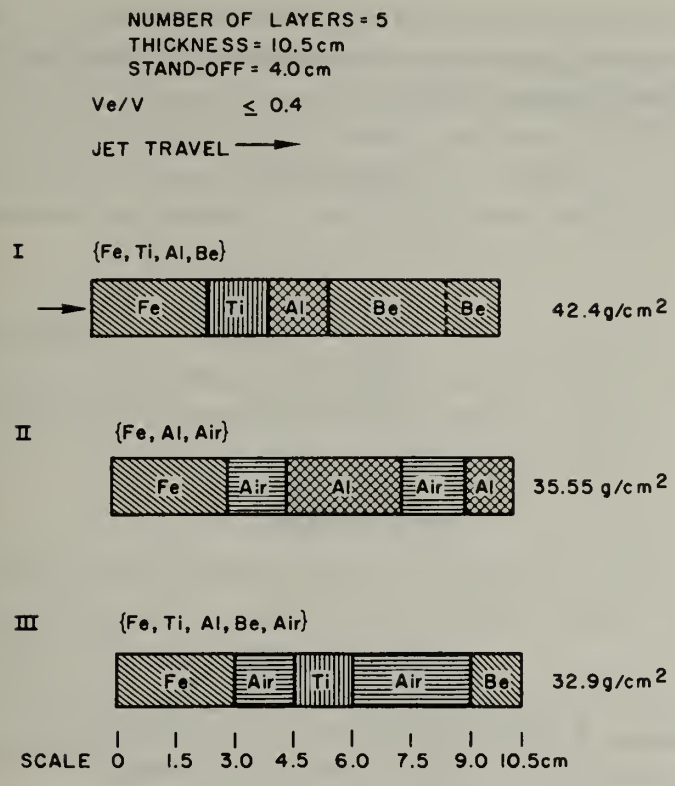


FIGURE 2. *Effect of component preselection upon minimum mass/area.*

Finally we illustrate in Table II the variation in the minimum mass per unit area with total armor ness. Again we assume a given shaped charge threat at a standoff distance of 4.0 cm and require the ratio of residual jet exit velocity to the initial striking velocity be less than or equal to 0.4. number of layers varies from 1 to 5 and the minimum layer thickness in all cases is equal to 1.5 cm.

TABLE II.

Minimum Mass/Area (g/cm^2) as a Function of Armor Thickness, T

N \ T	7.5 cm	9.0 cm	10.5 cm
1	58.50	70.20	47.40
2	48.63	47.25	43.65
3	46.80	40.56	39.60
4	46.80	39.66	38.30
5	46.80	36.93	32.90

INIMUM EXIT VELOCITY ARMOR

Another area of interest arises when one is confronted with the development of an armor con- tion subject to constraints imposed upon the total thickness and the mass per unit area. In this

case, one is usually interested in an inequality constraint upon the mass, i.e., the mass per unit ar must not exceed a given value. In addition, one is interested in minimizing the residual jet tip e velocity subject to the above constraints. That is

$$\underset{\{t_i, \rho_i\}}{\text{Minimize}} \; V_e/V$$

subject to

$$\sum_{j=1}^{N} t_j = T$$

$$\sum_{j=1}^{N} \rho_j t_j \leq M \tag{10}$$

and

$$\rho_j \in \{\rho^{(1)}, \ldots, \rho^{(k)}\}$$

for a given standoff distance Z. The expression for V_e/V is Equation (6).

The above problem can be cast as a dynamic program by defining:

$G_N(T, Z, M)$ as the minimum value of the ratio of the jet exit velocity to the initial striking velocit of a jet of a shaped charge, at standoff distance Z, which is directed against an N-layer armor of tot thickness T, and mass per unit area less than or equal to M.

The decision variables again are ρ_N and t_N and the choice of ρ_{N-1}, t_{N-1}, . . . must constitute a optimal policy. In this manner the following recursive relation is obtained:

$$G_N(T, Z, M) = \underset{t_n}{\text{Min}} \; \underset{\rho_N}{\text{Min}} \; \{g_N(t_N, \rho_N, Z) G_{N-1}(T - t_N, Z + t_N, M - \rho_N t_N)\}. \tag{11}$$

If we let the set (4) contain density values for metals only, we again note that if $\rho_i > \rho_j$, then $g_i < g$ By ordering the densities in decreasing value, $\rho^{(1)}, \rho^{(2)}, \ldots \rho^{(k)}$, G_1 can be defined as: Let r denote th smallest value of the superscripts $1, 2, \ldots, k$ for which $\rho^{(r)}$ satisfies $\rho^{(r)} T \leq M$ then

$$G_1(T, Z, M) = g_r(Z, T, \rho^{(r)}). \tag{12}$$

If there does not exist any such r for a given M, $G_1(T, Z, M)$ is defined to be infinite, signifying that ther does not exist any material, in (4), which satisfies the mass constraint.

The computation, utilizing Equations (11) and (12) can proceed as previously described. Again w note that $G_N(T, Z, M)$ is a discontinuous function of M; therefore, the work of Haymond is agai applicable. In particular, for a given value of M, there exists an armor configuration whose mass pe unit area is $M_A \leq M$ and which has a minimum value of V_e/V. The values of V_e/V and M_A will, in ger eral, be the same as for the corresponding solution of Equations (2)–(6), hence, the solution to the abov problem is not further described.

CONCLUSION

In the preceding we have described two formulations of problems whose solutions are useful in development of armor configurations for use in defense of shaped charge threats. The rationale eloped is for continuous jets; however, the resulting configurations are directly applicable in the elopment of such armor. The dynamic programming formulation provides useful information in ard to the sensitivity of the solution to change in thickness, number of layers, etc., during the rse of solution.

. ACKNOWLEDGMENT

The authors thank Dr. Frank Grubbs for his valuable assistance and encouragement in developing presenting this approach to problems of engineering design.

REFERENCES

Bellman, R., *Adaptive Control Processes* (Princeton University Press, 1961).

Birkhoff, G., D. P. McDougal, E. M. Pugh and G. I. Taylor, "Explosive with Lined Cavities," J. Appl. Phys. *19*, 563–582 (1948).

Haymond, R. E., "Discontinuities of Optimal Return in Dynamic Programming," J. Math. Anal. & Applic. *30*, 639–644 (1970).

ON MULTIPOPULATION QUEUING SYSTEMS WITH FIRST-COME FIRST-SERVED DISCIPLINE

K. Truemper

University of Texas at Dallas

ABSTRACT

Steady-state probabilities for a multipopulation single channel multiserver queuing system with first-come first-served queue discipline are established. Special cases lead to simplified formulas which sometimes can be evaluated using existing statistical and/or queuing tables. Optimization aspects are considered involving control of service rates. Several applications are presented.

INTRODUCTION

Real world queuing systems frequently receive input from several populations with differing val patterns and/or service requirements. If queue discipline and service rules are such that the iber of possible states for the system is small, one can often derive closed form solutions, or at t solve numerically, for state probabilities and distributions of interest. This is, for example, the case n priority rules are imposed. On the other hand, some queue disciplines, among them the first- e first-served (FCFS) rule, typically lead to a large number of different states, and only in special s, e.g. when all populations are infinite and arrivals are according to a Poisson process with rate r population i as in [1], can the number of states be reduced by arguments relying on the particular al and/or service pattern at hand.

Here a multipopulation queuing system of type $M/M/s$ with a large number of different states is duced and solved for the steady-state probabilities, using an approach that seems to be promising ther multipopulation queuing problems of the exponential type as well. Assumed are r different ilations. A customer of population i arrives according to a Poisson process with rate $\alpha_j\lambda_{i,t(i)}$ if the number of customers in the system, i.e., in the queue or being served, is j, and the number of omers of population i in the system is $t(i)$. Upon arrival, each customer enters a single channel ie with FCFS discipline, or goes immediately into one of s servers if possible. Service times are nentially distributed, and the service rate μ_j of any occupied server depends only on the total num- of customers in the system, hence is independent from the type of customer that is being served. It be seen below that the assumption of exponential interarrival and service times is essential for the tion method employed. This assumption clearly restricts the applicability of the model, though uently insight into the characteristics of a system can be still gained from the results given here if actual distributions can be approximated by exponential ones.

Examples of the system described above are the parts issue department of an automobile repair , and a long-distance calling system of a business with leased telephone lines. In the first (second) , populations are the car mechanics of the shop and all customers buying parts (employees of the rent departments of the company), and the servers are the parts issue clerks (telephone lines).

Applications for the model typically arise when services provided for customers of different populations are identical or very similar in nature, and when the impact of system changes on quality of service for a particular population is to be evaluated. For example, in the repair shop problem one might want to know the effect on average waiting time for the car mechanics as the number of parts issue clerks is varied. As will be seen below, the solution is of sufficient simplicity to allow closed-form expressions and usage of existing statistical distribution and/or queuing tables in special cases. Further, the results lend themselves to queuing system optimization, including considerations of opening and closing additional servers.

Few results have been published about multipopulation queuing systems with FCFS discipline. All such systems assume Poisson arrivals and exponential service times. Cohen [4] has developed waiting time results for a single-server system with two infinite populations and different service rates for each population. For the same system providing service to r infinite populations, Ancker and Gafarian [1] have established steady-state probabilities and waiting-time results. The case of a single-server system with one finite and one infinite population requiring identical service has been solved in [16, 17]; the first reference includes numerical approaches and a computer program for transient probabilities and system descriptors. Lastly, Kotiah and Slater [9] have developed expected queue length and waiting time for the two-server system handling customers of two infinite populations with different service requirements. The latter paper contains additional references of related queuing systems.

The subsequent section 2 presents the steady-state equations for the queuing system introduced above. Steady-state probabilities are then derived in section 3 using an approach which appears to be promising for other multipopulation queuing systems of the exponential type. Section 4 specializes the results of section 3 for particular queuing systems. Lastly, optimization problems of service rate control are discussed in section 5.

2. STEADY-STATE EQUATIONS

The following notation will be used throughout:

r = number of populations

s = number of servers

$K=[k(1), \ldots, k(r)]$ = vector representing the number of customers in the queue; $k(i)$ = number of customers of population i

$L=[l(1), \ldots, l(r)]$ = vector representing the number of customers in the servers analogous to K

$T=K+L$

$$m=\sum_{i=1}^{r} l(i)$$

$$n=\sum_{i=1}^{r} k(i)$$

$$q=\sum_{i=1}^{r} t(i)=m+n$$

$p_{K,L}^{ab\cdots x}$ = steady-state probability that $k(i)$ $(l(i))$ customers of population i are in the queue (being served), $i=1, 2, \ldots, r$, the particular queue order being indicated by $ab \ldots x$; each

letter of the ordered n-tuple $ab \ldots x$ represents an integer between 1 and r and denotes the type of customer in the corresponding queue position, where customers arrived in order $x, \ldots, b, a$.

$$_{K,L} = \sum_{ab \ldots x} p^{ab \ldots x}_{K,L}$$

$$p_T = \sum_{K+L=T} p_{K,L}$$

$_{t(i)}$ = arrival rate for population i if the total number of customers in the system is q, of which $t(i)$ are of population i.

$$\alpha_q,\ \lambda_{i,\,t(i)} \geq 0; \qquad 0 \leq t(i) \leq q;\ \forall i$$

$$_{i}\lambda_T = \sum_{i=1}^{r} \alpha_q \lambda_{i,\,t(i)}$$

μ_q = service rate of each occupied server if the total number of customers is q.

$$\mu_q > 0;\ \forall q > 0$$

For brevity of formulation, the rate μ_0, defined to be zero, will be used. Further, $K+i$ will denote ector $[k(1), \ldots, k(i-1), k(i)+1, k(i+1), \ldots, k(r)]$; if K is the null vector, we will write i id of $0+i$. Analogous definitions hold for $L+i$.

To assure existence of a steady-state solution, we will require for the remainder of this section, as is in section 3, that $\alpha_j = 0, \forall j > j_{\max}$, where $j_{\max}$ is nonnegative and finite. This restriction implies he maximum number of customers allowed in the system is bounded by $j_{\max}$, which is almost s the case in real world applications. The case of infinite populations is discussed in section 4. With the above definitions the steady-state difference equations for the queuing system are

$$(\alpha_m \lambda_L + m\mu_m) p_{0,L} = \sum_{i=1,\ l(i)\neq 0}^{r} \alpha_{m-1} \lambda_{i,\,l(i)-1} p_{0,\,L-i}$$

$$+ \sum_{j=1}^{r} (l(j)+1)\mu_{m+1} p_{0,\,L+j};\ 0 \leq m < s$$

$$(\alpha_s \lambda_L + s\mu_s) p_{0,L} = \sum_{i=1,\ l(i)\neq 0}^{r} \Big[\alpha_{s-1} \lambda_{i,\,l(i)-1} p_{0,\,L-i}$$

$$+ \sum_{j=1}^{r} (l(j)+1-\delta_{ij})\mu_{s+1} p_{i,\,L-i+j} \Big];\ m = s$$

$$(1c) \qquad (\alpha_{n+s}\lambda_T + s\mu_{n+s})p^{ab}_{K,L}{}^{\cdot\cdot x} = \alpha_{n+s-1}\lambda_{a,\,l(a)-1}p^{b\,\cdot\cdot}_{K-a,\,L}{}^{\cdot\cdot x}$$

$$+ \sum_{i=1,\ l(i)\neq 0}^{r} \sum_{j=1}^{r} (l(j)+1-\delta_{ij})\mu_{n+s+1}p^{ab\,\cdot\cdot i}_{K+i,\,L-i+j}{}^{\cdot\cdot xi};\ m=s;\ n>0$$

where δ_{ij} is the Kronecker Delta. In accordance with definitions by other authors (see, e.g., Morse [12, *p.* 68]), Equation (1c) is the queue equation while Equations (1a, b) are the boundary conditions.

3. STEADY-STATE PROBABILITIES

The task of finding a solution for (1a–c) consists of two parts; (a) finding a general solution for the queue equations, and (b) specializing the general solution so that the boundary conditions are satisfied. Locating a general solution for the queue equations of multipopulation queues is usually a difficult matter since the underlying coefficient matrix typically does not exhibit the triangular or almost triangular structure which so frequently occurs in single population queues. For example, in the system considered by Ancker and Gafarian [1] and mentioned in the Introduction, restatement of the equations resulted in a block-triangular coefficient matrix, each block involving r equations. The approach taken here to obtain the general relationship between probabilities of the queue equations is quite different from the one in [1], but is an extension of the methods employed by Morse [12] for single population queues. One first replaces each (sometimes complex) transition occurring in the queue equation by one or more elementary transitions, and estimates a separate relation for each such elementary transition. Using the queue equation, the elementary relations are then combined into the desired general relationship, which is then specialized to satisfy the boundary conditions. Main difficulty of this approach is the estimation of consistent elementary relations where consistency is defined as follows:

Generally the sequence of elementary transitions that move the system from one state S_1 to some other state S_2 is not unique. Let $P[S_i]$ be the probability that the system is in state S_i, $i=1, 2$. Using the elementary relations, one can compute $P[S_2]$ starting with $P[S_1]$ and following a particular sequence of elementary transitions leading from state S_1 to state S_2. The elementary relations are then said to be consistent if for any states S_1, S_2, $P[S_2]$ computed from $P[S_1]$ as outlined is the same regardless of the sequence of elementary transitions that lead from S_1 to S_2. If the number of states is finite, and if all states are intercommunicating, as is the case here, then consistency implies that for any state S the probability $P[S]$ can be uniquely computed from $P[S_0]$ where S_0 is some arbitrarily selected state, usually representing the empty system. Furthermore, the probabilities so computed must by construction satisfy the steady-state equations, hence are indeed the unique solution to these equations (see Cox and Miller [5, *p.* 185]).

For the problem at hand, we will need three elementary transitions concerning (1) reordering the sequence of customers in the queue, (2) addition of a customer to the queue, and (3) exchanging a customer between the head of the queue and a server position.

In view of the results for one finite and one infinite population [16, 17], we tentatively estimate that

$$(2a) \qquad p^{ab}_{K,L}{}^{\cdot\cdot x} = p^{\overline{ab}}_{K,L}{}^{\cdot\cdot \overline{x}};\ m=s;\ n>0$$

e $\overline{ab}\ldots\overline{x}$ is an arbitrary permutation of $ab\ldots x$. For addition of a customer to the queue, robabilities are to be related by

$$p_{K,L}^{ab\cdots x} = v_{a,K-a,L}\, p_{K-a,L}^{b\cdots x} .$$

ly. exchanging a customer between the head of the queue and a server position is to be described by

$$p_{K+i,L+j}^{ab\cdots xi} = w_{i,j,K+j,L+i}\, p_{K+j,L+i}^{ab\cdots xj} .$$

To determine the unknown functions of (2b, c) we first choose K and L in the queue equation (1c) so $p_{K,L}^{ab\cdots x}$ and $p_{K-a,L}^{b\cdots x}$ are positive while $p_{\bar{K},\bar{L}}$ is equal to zero for any $\bar{K}$, $\bar{L}$ corresponding to $n+s+1$ omers. Since $\alpha_j = 0$, $\forall j > j_{\max}$, at least one such equation exists, or the queue equation can liminated from consideration entirely. Under the restrictions imposed on K and L, we have λ_T equal to zero, and relations (2a–c) inserted into (1c) yield

$$s\mu_{n+s}\, v_{a,K-a,L}\, p_{K-a,L}^{b\cdots x} = \alpha_{n+s-1}\, \lambda_{a,t(a)-1}\, p_{K-a,L}^{b\cdots x}$$

$$v_{a,K-a,L} = \alpha_{n+s-1}\, \lambda_{a,t(a)-1} / (s\mu_{n+s}) .$$

neral, relations (2a–c) applied to (1c) result in

$$(\alpha_{n+s}\, \lambda_T + s\mu_{n+s})\, v_{a,K-a,L}\, p_{K-a,L}^{b\cdots x} = \alpha_{n+s-1}\, \lambda_{a,t(a)-1}\, p_{K-a,L}^{b\cdots x} + \sum_{i=1,\, l(i)\neq 0}^{r} \sum_{j=1}^{r} (l(j)+1-\delta_{ij})\mu_{n+s+1}\, w_{i,j,K+j,L}\, v_{j,K,L}\, v_{a,K-a,L}\, p_{K-a,L}^{b\cdots x} .$$

and L in Equation (5) be chosen so that $p_{K,L}^{ab\cdots x}$ is positive, and furthermore that the number of mers is one less than the number of customers in (3). Then $p_{K-a,L}^{b\cdots x}$ is also positive, and $v_{j,K,L}$ of either equal to zero or given by (4) with subscripts properly changed. Now $v_{j,K,L}$ is equal to zero $_{+s}\, \lambda_{j,t(j)}$ is equal to zero, and (4) with changed subscripts applies in both cases, i.e.,

$$v_{j,K,L} = \alpha_{n+s}\, \lambda_{j,t(j)} / (s\mu_{n+s+1}) .$$

the above conclusions, we derive from (5)

$$v_{a,K-a,L} = \alpha_{n+s-1}\, \lambda_{a,t(a)-1} \Big/ \Big[\alpha_{n+s}\, \lambda_T + s\mu_{n+s} - \sum_{i=1,\, l(i)\neq 0}^{r} \sum_{j=1}^{r} (l(j)+1-\delta_{ij})\mu_{n+s+1}\, w_{i,j,K+j,L}\, \alpha_{n+s}\, \lambda_{j,t(j)} / (s\mu_{n+s+1}) \Big] .$$

ote that $v_{a,K-a,L}$ has the same form as (4) if the $w_{i,j,K+j,L}$ of (7) are such that

$$\lambda_T - \sum_{i=1;\, l(i)\neq 0}^{r} \sum_{j=1}^{r} (l(j)+1-\delta_{ij})\, w_{i,j,K+j,L}\, \lambda_{j,t(j)} / s = 0 .$$

This is easily achieved if

$$w_{i,j,K+j,L} = l(i)/[l(j) + 1 - \delta_{ij}]. \tag{9}$$

Proceeding inductively, we see that (4) and (9) used for (2b) and (2c), respectively, indeed satisfy a queue equations. To ascertain that these relations are consistent is a straightforward matter, and wi be omitted. The next step is finding relations for equations (1a, b). The procedure is analogous to the on for the queue equation, leading inductively to the final result

$$p^{ab\ldots x}_{K,L} = \left\{ \prod_{i=1;\, l(i)\neq 0}^{r} \left[\prod_{j=0}^{l(i)-1} \lambda_{i,j} \right] / l(i)! \right\} \left\{ \left[\prod_{j=1}^{q} \alpha_{j-1}/\mu_j \right] / s^n \right\} p_{0,0}; T \neq 0, \tag{10}$$

which upon substitution into (1a–c) confirms the correctness of the intermediate steps.

Applying (10) to a particular system, one can either develop formulas for system descriptors, o at least evaluate such descriptors numerically, using

$$p_{K,L} = \sum_{\overline{ab\ldots x}} p^{\overline{ab}\ldots\overline{x}}_{K,L} = p^{ab\ldots x}_{K,L}\, n!/\prod_{i=1}^{r} k(i)! \tag{11}$$

and

$$p_{0,0} = \left[1 + \sum_{T\neq 0} p_T/p_{0,0} \right]^{-1} \tag{12}$$

where p_T is defined in section 2. For single-server systems, the above relations simplify to

$$p_T = q! \left\{ \prod_{i=1;\, l(i)\neq 0}^{r} \left[\prod_{j=0}^{l(i)-1} \lambda_{i,j} \right] / l(i)! \right\} \cdot \left\{ \prod_{j=1}^{q} \left[\alpha_{j-1}/\mu_j \right] \right\} p_{0,0};\ T \neq 0. \tag{13}$$

When utilizing the well-known relationships concerning system descriptors, for example $L = \lambda \cdot W$ (for a "last word," see Stidham [15]), one can compute the average arrival rate for population i, $\overline{\lambda}_i$, by

$$\overline{\lambda}_i = \sum_{K,L} l(i)\mu_q p_{K,L} = \sum_{T} \alpha_q \lambda_{i,l(i)} p_T \tag{14}$$

since average arrival and service rates for each population must be equal in statistical equilibrium. Here, we will omit further details for the general case since the relevant formulas are easily available from the literature (see, e.g., Morse [12]).

4. SPECIAL CASES

Several multipopulation queuing systems are special cases of the one described by (1a–c). We will indicate a few such systems below.

1. Unbounded Queue Length

One can view such a system as the limit of a sequence of finite state space systems, and under rather mild assumptions steady-state probabilities $p_{K,L}$ are given by (11) if $\sum_{K+L\neq 0} p_{K,L}/p_{0,0}$ is convergent (see Feller [7, p. 455] for details and further references; for an elementary discussion, see Cox and Smith [6, p. 44]).

2. Balking

If $\alpha_j \leq 1$, $\forall j \geq 0$, factor $(1-\alpha_j)$ may be viewed as the probability that an arriving customer
y population balks (i.e., refuses to join the system) if the total number of customers already in the
m is equal to j. In the general case the average rate of balking for population i is

$$\sum_T (1-\alpha_q)\lambda_{i,t(i)}p_T$$

$_{t(i)}$ is the arrival rate for that population when $t(i)$ customers of type i are already in the system.

3. Finite and Infinite Populations

In general, computation of $p_{0,0}$ of (12) involves a very large number of probabilities $p_{K,L}$ defined
1); similarly, evaluation of system descriptors typically requires extensive computational efforts.
ecial cases, however, computations can be drastically reduced by the use of tables for certain
tical distributions and/or queuing systems. As an example, we will consider a queuing system with
inite and one infinite population, noting that both applications cited in the Introduction may be
s type. Arrival rates are

$$\alpha_q\lambda_{1,t(1)} = \begin{cases} (h-t(1))\lambda_1; & 0 \leq t(1) \leq h \\ 0; & \text{else} \end{cases}$$

e finite population, which is assumed to be of size h, and

$$\alpha_q\lambda_{2,t(2)} = \lambda_2$$

e infinite population. Service rates are $\mu_j = \mu$, $\forall j > 0$, for any occupied server. Inserting these
into (10) and assuming $\lambda_2 < s\mu$, we get for $0 \leq x \leq \min\{h, s\}$

$P[x] = P$ [x finite population customers in the servers]

$$= \left\{ \left[1-\delta_{x,s} \right] \left[\binom{h}{x} \left(\frac{\lambda_1}{\mu}\right)^x \sum_{j=0}^{s-x-1} \left(\frac{\lambda_2}{\mu}\right)^j /j! \right] \right.$$

$$\left. + \left[\binom{s}{x} \left(\frac{\lambda_2}{\mu}\right)^{s-x} /s! \right] \left[\sum_{i=x}^{h} \binom{h}{i} \left(\frac{\lambda_1}{\mu}\right)^i i!/s^{i-x} \sum_{j=0}^{\infty} \binom{j+i-x}{i-x} \left(\frac{\lambda_2}{s\mu}\right)^j \right] \right\} p_{0,0}$$

$$= \{ [1-\delta_{x,s}]\, \beta_x\gamma_{x-s-1} + c(\zeta_x\eta_{s-x}) \} p_{0,0}$$

$$\beta_x = \binom{h}{x}\left(\frac{\lambda_1}{\mu}\right)^x$$

$$\gamma_{s-x-1} = \sum_{j=0}^{s-x-1} \left(\frac{\lambda_2}{\mu}\right)^j \Big/ j!$$

$$\zeta_x = \sum_{i=x}^{h} \binom{h}{i} [\lambda_1/(s\mu - \lambda_2)]^i i! \tag{18}$$

$$\eta_{s-x} = \binom{s}{x} \left(\frac{\lambda_2}{s\mu}\right)^{s-x} \left(1 - \frac{\lambda_2}{s\mu}\right)^x$$

$$c = s^s / \left[s! \left(1 - \frac{\lambda}{s\mu}\right) \right].$$

When $\lambda_2 \geq s\mu$, $P[x]/p_{0,0}$ becomes unbounded, and the system never attains statistical equilibrium. In view of the results for infinite population queues cited above, $\lambda_2 < s\mu$ is therefore necessary and sufficient for the existence of a steady-state solution, and $P[x]$ is given by (17) in that case.

In (18), both β_x and ζ_x can be computed from the finite queuing tables by Peck and Hazelwood [13] by formulas of Votaw and Peck [18], while γ_{s-x-1} and η_{s-x} are available from tables of the Poisson distribution (Molina [11]) and binomial distribution (for references, see Feller [7, p. 148]), respectively. Further, the binomial distribution may be approximated by the Poisson distribution, and both may be approximated by the normal distribution (Feller [7, p. 154, 188, 245]). The factor $s^s/s!$ of c can by Stirling's formula be estimated as $e^s/(2\pi s)^{1/2}$ (Feller [7, p. 52]).

From (17) the probability that the system is empty is

$$p_{0,0} = \left\{ \sum_{x=0}^{\min(h,s)}{}' [(1 - \delta_{x,s})\beta_x \gamma_{s-x-1} + c(\zeta_x \eta_{s-x})] \right\}^{-1} \tag{19}$$

For a single-server system, (19) reduces to

$$p_{0,0} = (1 - \lambda_2/\mu)\xi_0 \tag{20}$$

where ξ_0 is the probability that a single-server system for one finite population is empty when the arrival rate is given by (15) and the service rate is equal to $(\mu - \lambda_2)$. $p_{0,0}$ of Equation (20) is also obtained when p_T of (13) is inserted into (12), and is readily computed using the finite queuing tables [13].

In an analogous fashion, formulas for system descriptors can be developed and related to existing tables. We will omit details since equations (17) and (18) indicate the general approach. For single-server systems, the relevant formulas are provided in [16, 17]. Lastly, it should be noted that the above results are easily extended to a system with one finite and several infinite populations by defining λ_2 in (17)–(20) to be the sum of the arrival rates for the infinite populations.

5. OPTIMIZATION ASPECTS

The definitions of arrival and service rates in section 2 admit formulation of a variety of optimization problems involving state-dependent control of these rates. Computational difficulties lie in the fact that in general the control-dependent steady-state probabilities lead to criterion functions which

ıot be optimized by reasonable computational efforts. However, special cases lend themselves to solution. For example, if all populations are infinite, results for single population queuing systems pe $M/M/s$ (see, e.g., Meyer [10]) can often be adapted to the system at hand. When one finite and ral infinite populations provide input, the computational approach outlined in section 4 may be loyed.

In contrast to the arrival control problem for two classes of customers considered by Scott [14], vill explore a particular optimization problem involving addition and removal of a server, which is ecial service rate control problem. In the system of section 2, let one server be added when queue h is equal to $n_1 \geq 1$ and one arrival occurs. As long as queue length is greater than n_1, the additional er is active. However, when queue length is equal to n_1 and one server finishes service, that server osed. We will first develop the steady-state probabilities, then present an application.

For the modified system, equations (1a–c) hold when $n < n_1$. Then

$$) \quad (\alpha_{n+s}\lambda_T + s\mu_{n+s})\, p^{ab\cdots x}_{K,L} = \alpha_{n+s-1}\lambda_{a,t(a)-1} p^{b\cdots x}_{K-a,L} + \sum_{j=1}^{r} (l(j)+1)\,\mu_{n+s+1} p^{ab\cdots x}_{K,L+j};\ m=s;\ n=n_1$$

$$) \quad (\alpha_{n+s+1}\lambda_T + (s+1)\,\mu_{n+s+1})\, p^{ab\cdots x}_{K,L} = \sum_{i=1;\, l(i)\neq 0}^{r} \left[\alpha_{n+s}\lambda_{a,t(a)-1} p^{b\cdots xi}_{K-a+i,L-i} + \sum_{j=1}^{r} (l(j)+1-\delta_{ij})\,\mu_{n+s+2} p^{ab\cdots xi}_{K+i,L-i+j} \right];\ m=s+1;\ n=n_1$$

$$) \quad (\alpha_{n+s+1}\lambda_T + (s+1)\,\mu_{n+s+1})\, p^{ab\cdots x}_{K,L} = \alpha_{n+s}\lambda_{a,t(a)-1} p^{b\cdots x}_{K-a,L} + \sum_{i=1;\, l(i)\neq 0}^{r} \sum_{j=1}^{r} (l(j)+1-\delta_{ij})\,\mu_{n+s+2} p^{ab\cdots xi}_{K+i,L-i+j};\ m=s+1;\ n>n_1.$$

ı applying the solution approach of section 3, we have

$$p^{ab\cdots x}_{K,L} = \left\{ \prod_{i=1;\, l(i)\neq 0}^{r} \left[\prod_{j=0}^{l(i)-1} \lambda_{i,j} \right] \Big/ l(i)! \right\} \cdot \left\{ \left[\prod_{j=1}^{q} \alpha_{j-1}/\mu_j \right] \Big/ \left[s^{\min(n,\, n_1)} (s+1)^{\max(n-n_1,\, 0)} \right] \right\} p_{0,0};\ T \neq 0.$$

emaining formulas concerning $p_{K,L}, p_T$, etc., are those of section 3.

The above system occurs in certain models of multiprocessor computer systems. Suppose two ssors solve high priority problems generated by $r \geq 1$ populations as well as low priority problems ıble in unlimited number (e.g., low priority processing time is sold to a computer network). Each ssor is assigned to a separate block of fast memory, and has access to either class of problems. system may be operated under any one of many possible priority rules. We will consider two rules

of the preemptive-resume priority type. Under rule 1, high priority jobs are handled FCFS and displace any low priority job in either processor upon arrival. Depending on arrival and service rates for high priority customers, this rule may lead to frequent "swaps" (removal of a job from fast memory and subsequent reinsertion) of low priority jobs. Rule 2 has one processor dedicated to low priority jobs unless the number of high priority jobs in the system exceeds n_1+1 where $n_1 \geq 1$. Selection of n_1 obviously affects the average rate of swaps as well as the expected waiting time for high priority jobs. Bell [3] has considered a related FCFS single-server system with two infinite populations. Besides cost of waiting for high priority customers, a penalty cost is incurred whenever the server elects to serve a high priority customer not at the head of the queue. Additional references concerning the single population case are also provided. Queuing models by Avi-Itzhak and Heyman [2] for single processor multiprogramming computer systems very well complement the model described above on certain multiprocessor computer system applications where processors do not interact. In such systems one can decompose scheduling decision problems into two parts to obtain approximate solutions. First the priorities problem is examined under simplifying assumptions for processing rules; the above model, as well as many others (see Jaiswal [8]), are potentially applicable. In part two, the concepts developed by Avi-Itzhak and Heyman are then employed to obtain detailed results for each priority class, utilizing system descriptor results of the priorities model.

Returning to the application at hand, we consider two FCFS single channel queuing systems with input from the high priority job populations. We will suppose that interarrival and service times satisfy the assumptions of the system considered in section 2. Under rule 1, we have a two-server system, so the steady-state probabilities are given by (10) and (11), and will be denoted by $p_{K,L}(0)$. Rule 2 leads to a single-server system with server addition (removal) when queue length is equal to n_1 the number of busy servers is one (two), and an arrival (service completion) occurs. Note that the server that is added is not necessarily the one that is removed. However, either processor can handle high or low priority jobs, so for our purposes it is irrelevant which type of job is assigned to a particular processor. Hence steady-state probabilities, denoted by $p_{K,L}(n_1)$, are given by (21) and (11). Utilizing $L=\lambda \cdot W$, expected steady-state waiting times for high priority jobs, $W(0)$ and $W(n_1)$ under rules 1 and 2, respectively, can be computed. In analogous notation, the steady-state rates for swaps are by elementary arguments

$$e(0)=\mu\left[\sum_{i=1}^{r} p_{0,i}(0)+2\sum_{L:\,m=2} p_{0,L}(0)\right]$$

$$e(n_1)=\mu\left[\sum_{i=1}^{r} p_{0,i}(n_1)+2\sum_{K:\,n=n_1;\,L:\,m=2} p_{K,L}(n_1)\right] \tag{22}$$

where μ is the service rate for high priority jobs.

Applying relevant cost/profit relationships to the above descriptors, the optimal rule is easily selected. For particularly simple systems, closed form expressions for the descriptors can sometimes be obtained, and minimization of the total cost function with respect to $n_1 \geq 0$ (where $n_1=0$ denotes rule 1) is straightforward.

SUMMARY

This paper establishes steady-state probabilities for a class of FCFS multipopulation queuing ems, assuming exponential interarrival and service times. The solution approach is an extension ethods employed by Morse [12] for single population queues, and appears promising for other ipopulation queues of the exponential type.

Future research on the topic of FCFS multipopulation queues seems desirable. In particular, lts for systems with nonexponential interarrival and service time distributions would be valuable. ther point should also be mentioned. It may be very difficult to obtain exact formulas for certain S systems, or the formulas may involve extensive computational efforts for evaluation. Simple and -to-apply approximation formulas would be of practical value, provided reasonable bounds can be blished for the errors introduced by such approximations.

REFERENCES

Ancker, C. J. and A. V. Gafarian, "Queuing with Multiple Poisson Inputs and Exponential Service Times," Operations Research *9*, 321–327 (1961).

Avi-Itzhak, B. and D. P. Heyman, "Approximate Queing Models for Multiprogramming Computer Systems," Operations Research *21*, 1212–1230 (1973).

Bell, C. E., "Efficient Operation of Optional-Priority Queuing Systems," Operations Research *21*, 777–786 (1973).

Cohen, J. W., "Certain Delay Problems for a Full Availability Trunk Group Loaded by Two Traffic Sources," Communication News *16*, 105–113 (1957).

Cox, D. R. and H. D. Miller, *The Theory of Stochastic Processes* (Methuen and Company, London, 1965).

Cox, D. R. and W. L. Smith, *Queues* (Chapman and Hall, London, 1961).

Feller, W., *An Introduction to Probability Theory and Its Applications—Volume I*, (Wiley and Sons, New York, 1968) 3rd Edition.

Jaiswal, N. K., *Priority Queues* (Academic Press, New York, 1968).

Kotiah, T. C. T. and N. B. Slater, "On Two-Server Poisson Queues with Two Types of Customers," Operations Research *21*, 507–603 (1973).

Meyer, K. H. F., *Wartesysteme mit Variabler Bearbeitungsrate* (Springer Verlag, Berlin, 1971).

Molina, E. C., *Poissons's Exponential Binomial Limit* (Van Nostrand, New York, 1942).

Morse, P. M., *Queues, Inventories and Maintenance* (Wiley and Sons, New York, 1958).

Peck, L. G. and R. N. Hazelwood, *Finite Queuing Tables* (Wiley and Sons, New York, 1958).

Scott, M., "Queuing with Control on the Arrival of Certain Type of Customers," J. Canadian Operational Research Soc. *8*, 75–86 (1970).

Stidham, S., "A Last Word on $L = \lambda \cdot W$," Operations Research 22, 417–421 (1974).

Truemper, K., "Single-Channel Single-Server Queues with Input from Both a Finite and an Infinite Population," Master's Thesis, University of Iowa (1969).

Truemper, K., "Queuing with Poisson Arrivals from Both a Finite and Infinite Population," AIIE Transactions *4*, 223–227 (1972).

Votaw, D. F. and L. G. Peck, "Remarks on Finite Queuing Tables," Operations Research *16*, 1084–1086 (1968).

INFORMATION FOR CONTRIBUTORS

he NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of ific information in logistics and will publish research and expository papers, including those tain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve ficiency and effectiveness of logistics operations.

lanuscripts and other items for publication should be sent to The Managing Editor, NAVAL ARCH LOGISTICS QUARTERLY, Office of Naval Research, Arlington, Va. 22217. manuscript which is considered to be suitable material for the QUARTERLY is sent to one re referees.

lanuscripts submitted for publication should be typewritten, double-spaced, and the author l retain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted he original.

short abstract (not over 400 words) should accompany each manuscript. This will appear head of the published paper in the QUARTERLY.

here is no authorization for compensation to authors for papers which have been accepted blication. Authors will receive 250 reprints of their published papers.

eaders are invited to submit to the Managing Editor items of general interest in the field istics, for possible publication in the NEWS AND MEMORANDA or NOTES sections QUARTERLY.

NAVAL RESEARCH
LOGISTICS
QUARTERLY

SEPTEMBE
VOL. 23, N
NAVSO P-

CONTENTS

ARTICLES

A Survey of Maintenance Models: The Control and Surveillance of Deteriorating Systems — W. P. PIERSKALL, J. A. VOELKE

Markovian Deterioration with Uncertain Information—A More General Model — D. ROSENFIEL

An Algorithm for the Segregated Storage Problem — A. W. NEE, M. R. RA

Optimal Facility Location under Random Demand with General Cost Structure — V. BALACHANDR, S. J

Some Comparative & Design Aspects of Fixed Cycle Production Systems — A. L. SOYST, D. I. TOO

Determining Adjacent Vertices on Assignment Polytopes — P. G. MCKEOW

Improved Convergence Rate Results for a Class of Exponential Penalty Functions — F. H. MURPH

Estimation of Strategies in a Markov Game — J. A. FILA

Flowshop Sequencing Problem with Ordered Processing Time Matrices: A General Case — M. L. SMIT, S. S. PANWALKA, R. A. DUDE

Mean Square Invariant Forecasters for the Weibull Distribution — J. TIAGO DE OLIVEI, S. B. LITTAU

A Comparison of Several Attribute Sampling Plans — A. E. GELFA

Suboptimal Decision Rule for Attacking Targets of Opportunity — T. KI

Occupational Structure in the Military and Civilian Sectors of the Economy — S. E. HAB

Armor Configurations via Dynamic Programming — R. E. SHEA, A. L. ARBUCKL, V. B. KUCH

On Multipopulation Queuing Systems with First-Come First-Served Discipline — K. TRUEMPE

OFFICE OF NAVAL RESEARCH
Arlington, Va. 22217